The Use of Non-volatile a-Si:H Memory Devices for Synaptic Weight Storage in Artificial Neural Networks

Andrew J. Holmes

Thesis submitted for the degree of
Doctor of Philosophy
The University of Edinburgh
January 1995

Abstract

This thesis describes the development of an ANN chip in which a-Si:H resistors are integrated with CMOS circuitry. This eliminates the need for external refresh or neuron circuitry required by ANN designs based on dynamic storage techniques. The a-Si:H memory technology was developed in collaboration with Dundee University and is in effect a programmable, non-volatile, semiconductor resistor. The device consists of a thin 1000Å layer of a-Si:H sandwiched between vanadium and chromium electrodes.

During the project a total of three test chips were designed and fabricated. The first chip was used to investigate the fabrication of memory devices on the surface of a CMOS wafer: previously all the test devices had been constructed on glass slides. Results from this chip showed that it was possible to fabricate programmable a-Si:H resistors on a CMOS chip. The second chip contained five different synapse designs all of which used the a-Si:H resistor as the memory element. The best of these was then used in the construction of the final ANN chip. This chip contained an 8 x 8 array of synapses and digital addressing, and required minimal support circuitry.

Conclusions are drawn both about the performance of the a-Si:H memory device and the alternative approaches to non-volatile storage in ANN chips, and recommendations are made for future work in this area.

Declaration

I declare that this thesis has been completed by myself and that, except where indicated to the contrary, the research documented in this thesis is entirely my own.

Andrew J. Holmes

Acknowledgements

There are a number of people to whom I am deeply indebted, for their help, advice and support during the course of this research.

- My academic supervisors, Alan Murray and Martin Reekie.
- The Amorphous Materials Group at Edinburgh - Janos Hajto, Tony Snell and Alan Owen for both assisting with the testing of the a-Si:H memory devices, as well as providing laboratory space and test equipment.
- Tony Reeder and Ian Thomas at BT for providing a case sponsorship as well as vacational employment prior to the official start of the project.
- The Amorphous Materials group at Dundee University - Rod Gibson, Merv Rose and PG LeComber for fabricating the memory devices.
- The staff at ES2, particularly Dave Easey, Graeme Cairns and Steve Barker for ensuring the smooth transition from GDSII file to completed wafer.
- Bill Seddon at Rutherford Appleton Laboratories for dealing with both ES2 and the mask manufacturers, Compugraphics.
- Alec Ruthven from the EMF for patiently bonding "yet another wafer segment".
- I must also extend my thanks to my fellow students, for making the last three years most enjoyable and rewarding. Particularly the hardware diehards - Steve Churcher, Donald Baxter, Alister Hamilton, Robin Woodburn, Geoff Jackson, Dwayne Burns and Andy Myles.
- Finally, I would like to acknowledge the contributions of the Science and Engineering Research Council, and BT, for personal financial support, and for technical funding.

Table of Contents

Chapter 1

Introduction and Thesis Overview

1.1. Artificial neural networks

In a conventional digital computer the processing is performed by a single, complex element, the microprocessor. However, there are many tasks, such as speech and vision processing, that are routinely performed by humans which these digital computers cannot perform. The "architecture" adopted by the human brain is radically different from that of a conventional digital computer: instead of one complex processing unit the brain uses a huge number of simple processing cells that are highly interconnected. This architecture based on parallel distributed processing is the basis of artificial neural network (ANN) research.

In an ANN only the most basic aspects of the complex bio-chemistry of a real biological nervous system are modelled. The result is a highly abstracted, "engineer friendly", model of a biological neural network, as shown in figure 1.1.

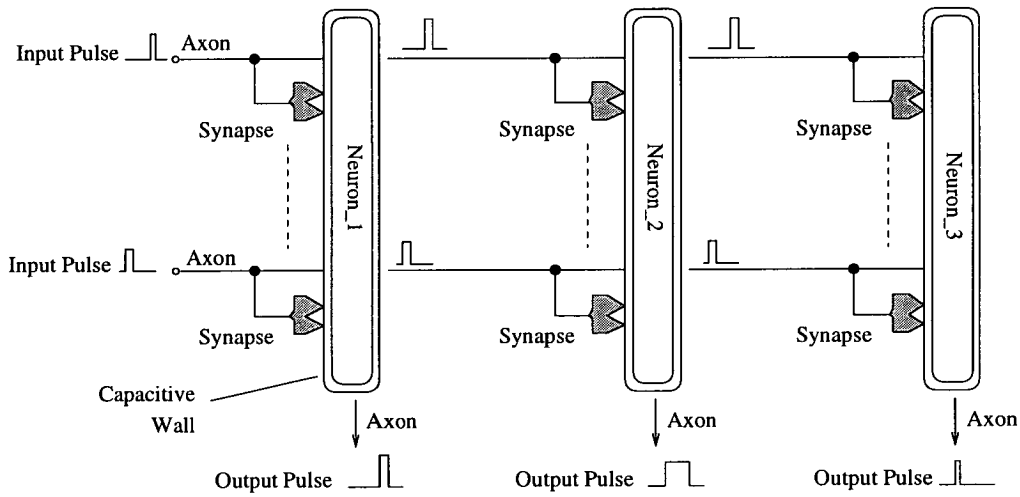


Figure 1.1 - Model of a biological neural network

The operation of this simple network can be summarised as follows:

• Input and output signals are encoded as trains of pulses: the higher the activity level the higher the pulse firing rate.

• The basic processing element is the neuron. This has a capacitive wall which "integrates" the various input current pulses and then outputs a voltage pulse once a given threshold level has been exceeded.

• The neuron's output pulses, which are transmitted along interconnections called axons, are coupled to other neurons through junctions called synapses.

• The strength, or weight, of the coupling provided by the synapse determines how much effect the output pulse from one neuron has on the input level of another. This synaptic weight, which can be either inhibitory or excitatory, can be thought of as a +/- multiplier term.

At the present time the bulk of neural network research is done using algorithms that run on general purpose digital computers; Lippmann provides a good introduction to some of the different network topologies and training algorithms that are used[1]. However, neural network architectures, by their nature - a distributed network of simple processing elements - lend themselves to analogue VLSI implementation. Analogue implementations promise significantly smaller chip area than their digital counterparts, both for circuitry and wire area. Analogue ANN chips have been used in a variety of applications including recognition of handwritten characters[2] and classifying heart arrhythmias[3].

Since 1987 the Neural Network Group at Edinburgh University has been steadily developing a series of analogue ANN chips, culminating in the EPSILON chip, which contained a 120 x 30 array of synapses[4]. On the EPSILON chip the synaptic weights are stored dynamically as voltages on capacitors[5].

Charge leakage means that the dynamic storage scheme used on EPSILON requires external refresh circuitry. A number of schemes have been proposed to increase the hold time of these weight capacitors by periodic comparison with a staircase waveform[6-11]. However, if a single standalone neural chip is required then some form of on-chip non-volatile weight storage is required. Such a chip could then act almost as an "intelligent" analogue to digital convertor, taking analogue signals in and then performing some classification determined by the programmed synaptic weights.

1.2. The amorphous silicon analogue memory

During research into the switching properties of thin films of amorphous silicon, researchers at Dundee and Edinburgh Universities observed a non-volatile, analogue switching behaviour in one of the structures. This device consisted of a 0.1 μm thick layer of p^+ amorphous silicon between vanadium and chromium electrodes, as shown in figure 1.2.

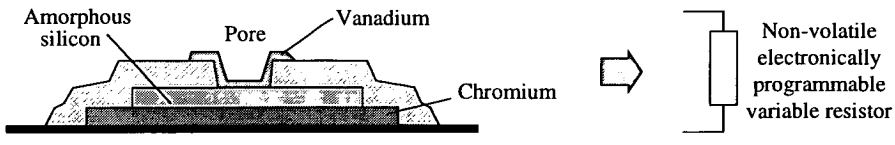


Figure 1.2 - The amorphous silicon resistor

This non-volatile resistor can be programmed to a resistance of between 1 kΩ and 1 MΩ with an accuracy of 5%[12] and devices have been shown to be stable, under conditions of zero bias, for up to 4 years[13].

1.3. Project Goal

The aim of the project described in this thesis was to replace the capacitor used for synaptic weight storage on the EPSILON chip with a non-volatile, a-Si:H resistor, as illustrated in figure 1.3. This project aim is intended to satisfy two goals: firstly to produce a a standalone ANN chip that does not require external refresh circuitry and secondly, to investigate the operation of the memory device in a practical application. In order to restrict the scope of the research it does not include investigating different neural architectures, beyond that used on EPSILON, or the mechanisms underlying the operation of the memory device itself.

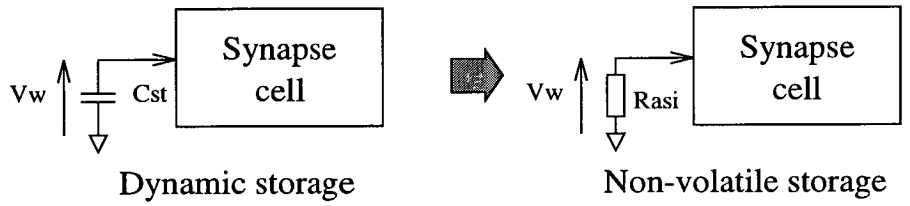


Figure 1.3 - Project aim

During the course of the project three test chips were designed and fabricated. Thus the thesis has been divided into four main sections, a review and then three chapters of experimental results, one on each of the chips:

Chapter 2		-	A review of non-volatile synaptic weight storage techniques
Chapter 3	Chip1	-	Integrating the a-Si:H resistor with CMOS technology
Chapter 4	Chip2	-	Synapse test circuits based on a-Si:H resistors
Chapter 5	Chip3	-	A neural network chip based on an 8x8 array of a-Si:H synapses

Chapter 2 contains a review of existing non-volatile storage techniques and development of the a-Si:H memory device.

Chapter 2

A Review of Non-volatile Synaptic Weight Storage

2.1. Introduction

This chapter contains a review of various techniques used for the non-volatile storage of synaptic weights in artificial neural networks.

Hardware implementations of artificial neural networks have used a wide variety of different technologies to perform the function of synaptic weight storage. Whilst some of these are unique to the neural network community, the majority are the direct descendants of their digital predecessors. As the development of digital memory devices may also hold some clues as to possible future trends in non-volatile weight storage this chapter will commence with a brief review of non-volatile, digital memory technologies[14]. The section that follows that will then look at synaptic weight storage, with particular emphasis on non-volatile, analogue storage techniques. The final section will introduce the a–Si:H analogue non-volatile memory, which has been developed jointly by Dundee and Edinburgh Universities.

This chapter can thus be divided into three main sections:

(i) Digital non-volatile storage

(ii) Non-volatile synaptic weight storage

(iii) The a-Si:H analogue memory device

2.2. Digital non-volatile storage

With the advent of digital computers in the 1960s there was a need for non-volatile memory chips to store the computer's "bootstrap" microcode, which instructs the computer to load the operating system from hard disk. In the early chips the patterns to be stored were defined by a customised mask, which had to be completed before chip fabrication. Even today this type of Read Only Memory (ROM) chip is still cost effective for high volume, hardwired applications. However, any alteration to the stored microcode requires the generation of a new mask set, which is both expensive and time consuming.

The limitations of these ROM chips provided the impetus for research into non-volatile memories chips which could be programmed electronically. The "memory" elements in these electrically programmable ROMs belong to one of two classes:

i) Devices in which a reversible change is induced electrically. This includes technologies such as floating-gate which will be discussed in section 2.1 .

ii) Fusible links which are irreversibly changed by an electrical pulse. Fuse based technologies will be discussed briefly in section 2.2.

Another technology that occupies a significant portion of the digital, non-volatile memory market is battery backed static RAM[15]. Modern devices have battery lifetimes of up to 5 years and and are expected to provide the memory capability in the new generation of "smart" credit cards[16]. As it is an "off-chip" method of achieving non-volatility it is technically outwith the scope of the following review.

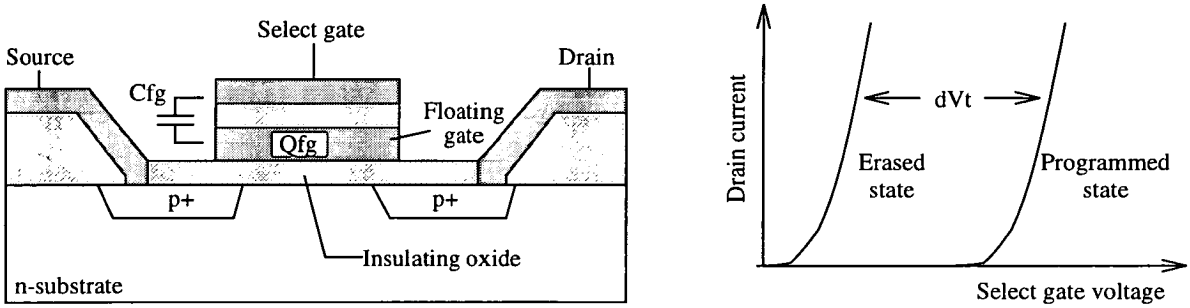
2.2.1. Reversible (non-fuse based) memory devices

This section describes the three electrically programmable memory technologies in which the state change is reversible. They are listed below in the order in which they will be discussed:

- Floating Gate - EPROM, EEPROM and Flash
- Silicon Nitride - MNOS and SONOS
- Ferroelectrics - FNVRAM

2.2.1.1. Floating gate storage

The first description of a programmable, non-volatile, semiconductor memory was published by Kahng and Sze in 1967[17]. By storing charge on a "floating" metal gate, placed above a conventional MOSFET as shown in figure 2.1(a), they recorded a bistable memory action.



(a) Cross-section (b) Operation

Figure 2.1 - The original floating gate memory.

The charge (Q_{fg}) stored on the floating gate alters the effective threshold voltage needed to turn the transistor on by an amount $dV_t = Q_{fg}/C_{fg}$, as shown in figure 2.1(b).

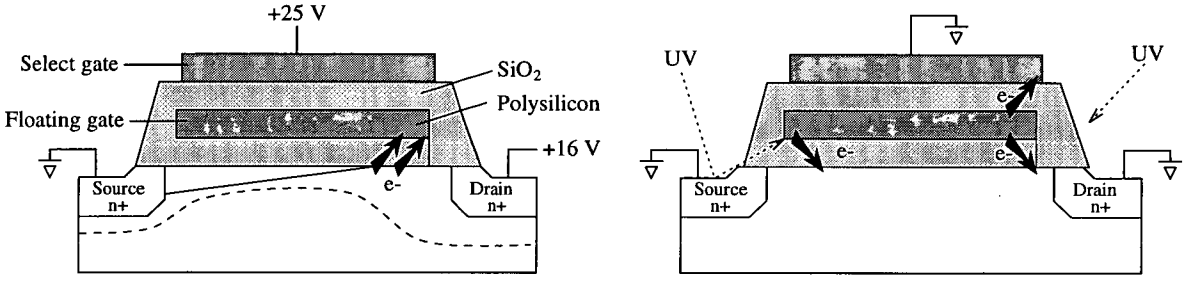
By applying a high field to the select gate, electrons are forced onto the "floating" gate causing the charge that is stored on it to increase. After the high field is removed the

charge remains, trapped by the surrounding insulator.

2.2.1.2. EPROM

The first commercially available Electrically Programmable ROM (EPROM) was produced by Intel in 1970[18]. The technology used was christened FAMOS - Floating Gate Avalanche Injection MOS.

The floating gate is charged via the avalanche injection of hot-electrons from the drain[19]: hot-electrons, so called because they have been accelerated by the high voltage applied to the EPROM drain, are pulled towards the floating gate by the high positive voltage on the select gate, as shown in figure 2.2(a).



(a) EPROM programming (b) EPROM erase

Figure 2.2 - EPROM programming and erase operations

The cell is erased using UV-light which gives the stored electrons enough energy to surmount the energy barrier between the floating gate and the insulator surrounding it.

2.2.1.3. EEPROM

The obvious disadvantage of EPROM technology is the need to place the chip in a special UV-eraser prior to reprogramming. In the late 1970s there was huge competition to produce a chip which could be erased, as well as programmed, electrically.

The first commercial Electrically Eraseable and Programmable ROM (EEPROM) chip was produced by Intel in 1980[20]. In their Floating-Gate Tunnel-Oxide (FLOTOX) device a very thin (200 Å) oxide between the floating polysilicon gate and an N+ diffusion region allowed both programming and erasing to be accomplished by electron tunnelling, as shown in figure 2.3.

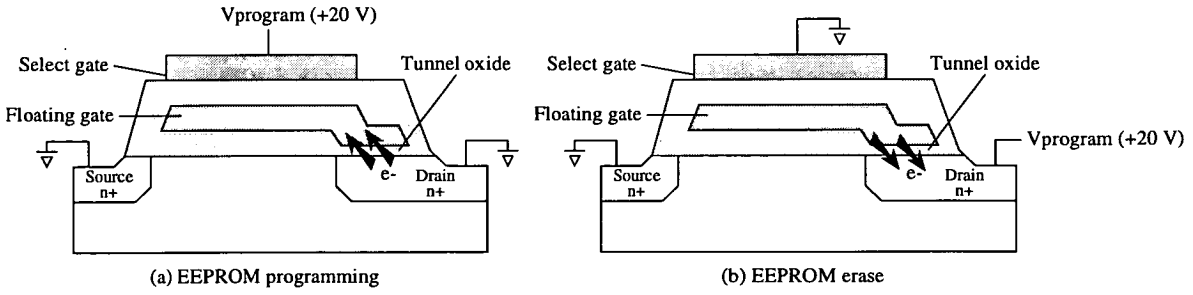


Figure 2.3 - EEPROM programming and erase operations

To program an EEPROM cell the drain is grounded and a positive programming pulse applied to the select gate, so attracting electrons onto the floating gate. Conversely, the cell is erased by grounding the gate and applying a positive programming pulse to the drain.

Unlike an EPROM cell a FLOTOX EEPROM cell contains two transistors, one for addressing and the other as a memory store, as shown in figure 2.4(a). However, by using specialised processing techniques, it is possible to make these EEPROM cells extremely compact. The cell shown in figure 2.4(b), which is taken from a 1 Mbit chip[21], uses a triple polysilicon process resulting in a cell that is only 3.8μm x 8μm.

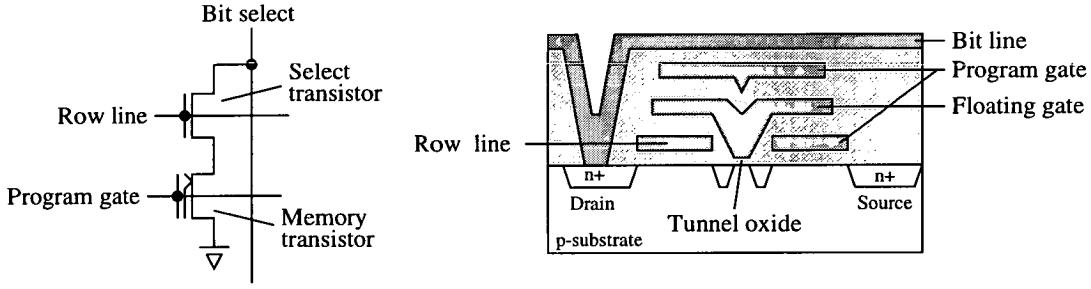


Figure 2.4 - A modern EEPROM cell

Other modern EEPROM designs are based on similarly specialised processing techniques; one example is the textured polysilicon cell[22] shown in figure 2.5.

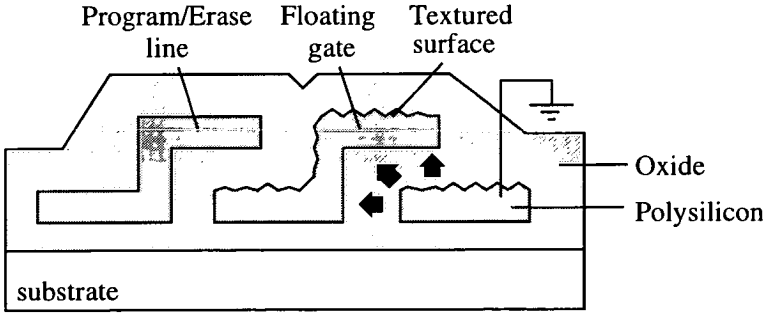


Figure 2.5 - Textured EEPROM cross-section

The fine texture at the polysilicon surface increases the local electric field during tunnelling by a factor of 3 to 5. This increased local field allows much thicker oxides, 600 to 1000 Å, to be used in the device construction, without the need for higher programming voltages. This thicker oxide means that failure due to oxide breakdown is less of a problem than for FLOTOX cells.

2.2.1.4. Flash EEPROM

Around 1984 the combination of hot electron programming and tunnel erase was rediscovered as a means of achieving single transistor EEPROMs[23]. However, unlike "full-featured" EEPROM chips these cannot be erased by bytes but must all be erased at once, resulting in the name "Flash" erase EEPROM.

The huge cost reduction that accompanies a single transistor based memory chip both in terms of size and decoder complexity has made Flash an alternative to EPROM for many applications, such as storing boot-up microcode[24]. The low cost also makes flash a viable alternative to magnetic media in some applications. Hitachi are developing the world's smallest camcorder based on 256 Mbit Flash chips. By using video compression they hope to store 30 minutes of digitised video in 400 Mbyte of flash memory[25].

2.2.1.5. Modern EEPROM support circuitry

Floating gate devices are programmed using relatively high, 15 V to 20 V, voltage pulses. In early chips these pulses had to be generated by external circuitry. Modern EEPROM chips now include on-board charge pump circuitry that allow them to be operated from a single +5 V supply. Catalyst have even developed an EEPROM which can generate the 18 V needed for programming from a 3 V battery power supply[26].

The new generation of Flash memories [27] also contain on-board analogue programmer circuitry to generate programming pulses of the correct shape. This results in devices with programming times as fast as 10μs/byte[28].

2.2.1.6. Silicon Nitride technologies - MNOS and SONOS

Non-volatile memory devices based on Metal-Nitride-Oxide-Silicon[29] structures have been around for almost as long as oxide based floating gate devices. An MNOS cell consists of a 20Å thermal SiO_2 gate dielectric and several hundred Angstroms of deposited silicon nitride, as shown in figure 2.6 (a).

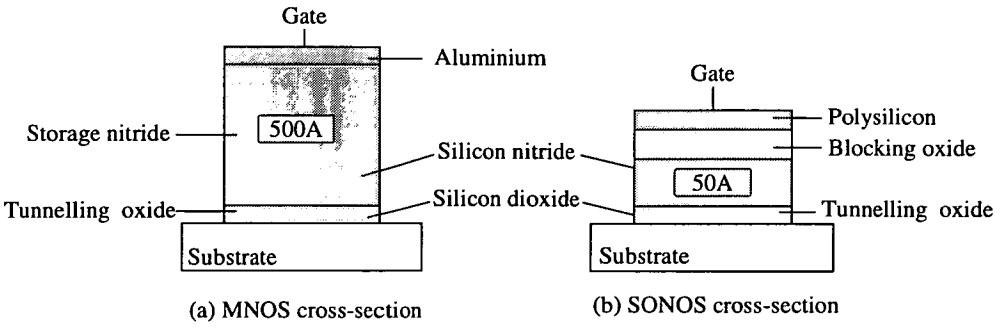


Figure 2.6 - MNOS and SONOS Transistor

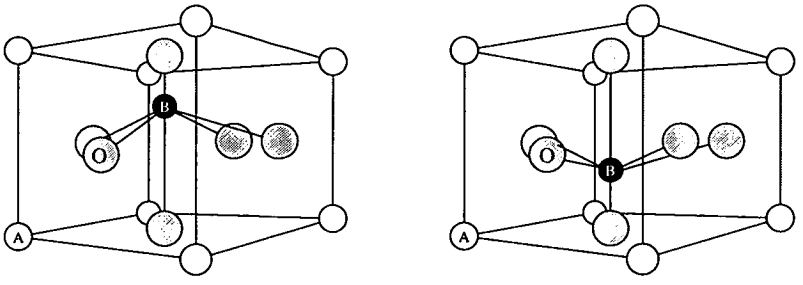
In an MNOS device the charge is not stored on a floating gate, as in EEPROM, but in discrete traps in the bulk of the nitride. During programming the charge transfer occurs over the large area of the channel region, unlike conventional floating gate where charge transfer occurs over a small area removed from the channel region. This means that there is less chance of device failure due to defects in the dielectric.

This raises the question as to why MNOS has always been a relatively minor player in field of high-density, non-volatile storage, when compared with FLOTOX devices. One reason is that, when EPROM manufacturers were considering an upgrade to an EEPROM technology, they had a choice of MNOS, which requires an ultra thin oxide and a high quality nitride, or FLOTOX where the processing is a simple variant on the standard high quality oxide furnace cycle used for EPROM devices[22].

In the 1980s an additional silicon gate was added to the MNOS device resulting in SNOS[30] and also an additional oxide to give SONOS devices. The use of the additional blocking oxide in a SONOS device allows the dielectric sandwich to be scaled to dimensions compatible with 5 - 10 V technology[31]. The reason for recent renewed interest in MNOS technology is that it is being promoted by two companies, Simtek in the US[30] and Hitachi in Japan[32].

2.2.1.7. Ferroelectrics

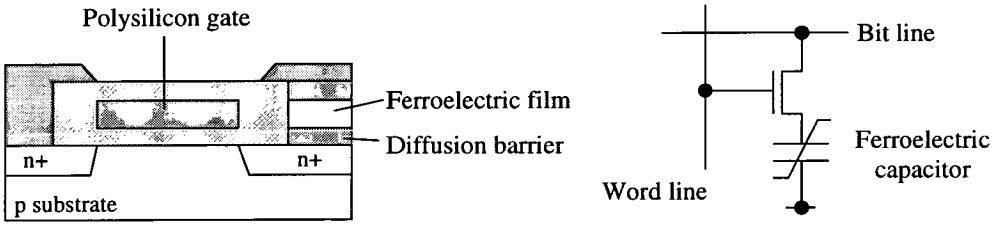
The most promising new area of non-volatile memory research is based on ferroelectric materials[33-35]. A ferroelectric film has a highly non-linear dielectric that retains the charge after an external voltage has been applied. This charge retention results from a net ionic displacement in the unit cells which have two stable states, a polarity of +0 or -1. Figure 2.7 shows a unit cell in the two different states.



(a) Polarity = +0 (b) Polarity = -1

Figure 2.7 - A Perovskite $AB0_3$ ferroelectric unit cell

The original devices in the 1960s suffered from problems of limited cycling and "read disturbance", where the read signal actually alters the stored memory state. In 1987 a new concept was introduced whereby a ferroelectric was used as the dielectric in a DRAM cell, as shown in figure 2.8.



(a) Cross-section (b) Cell Schematic

Figure 2.8 - A ferroelectric DRAM cell

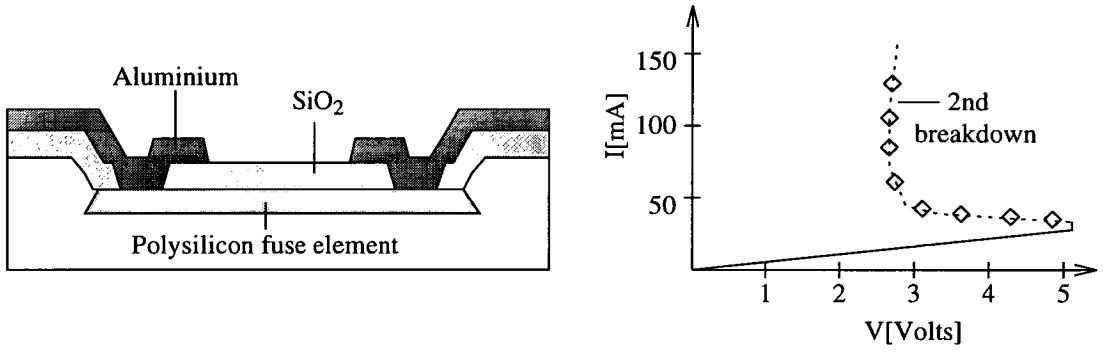
As the DRAM cell is continually refreshed the problem of read disturbance is eliminated, and because the dielectric is ferroelectric, the cell retains the stored charge when power is removed. The cell size and speed in this new technology, called Ferroelectric Non-Volatile RAM (FNVRAM), are expected to be comparable with today's conventional DRAM technology[27].

2.2.2. Fuses and antifuses

Fusible link technologies are mainly associated with Programmable-once ROM (PROM) devices. However, there is an increasing use of antifuses - initial high resistance transformed to a low resistance after programming - for field programmable gate arrays (FPGAs). This section covers firstly, conventional polysilicon fuses, and then two antifuse technologies. The switching characteristics of the silicon based devices are highlighted as they are close relatives of the amorphous silicon analogue memory.

2.2.2.1. Polysilicon fuses

Polysilicon resistors are used to provide the fusible links used in bipolar PROMs. The resistors can be constructed using either a horizontal[36] structure, as in figure 2.9(a), or a vertical sandwich structure[37]. The device is programmed by applying a high voltage pulse that causes the resistor to go open-circuit.



(a) Cross section (b) Programming characteristic

Figure 2.9 - Cross section and programming characteristic for a polysilicon fuse

During programming the device does not go open-circuit immediately: the large dissipated power causes contraction of the current to a filament, at some point where the film temperature has exceeded a critical value. This second breakdown state[36] does not always cause an open circuit and a further increase in pulse height may be required to fuse the device. The characteristic shown in figure 2.9(b) was built up by recording the steady state values of V and I during programming pulses of increasing height.

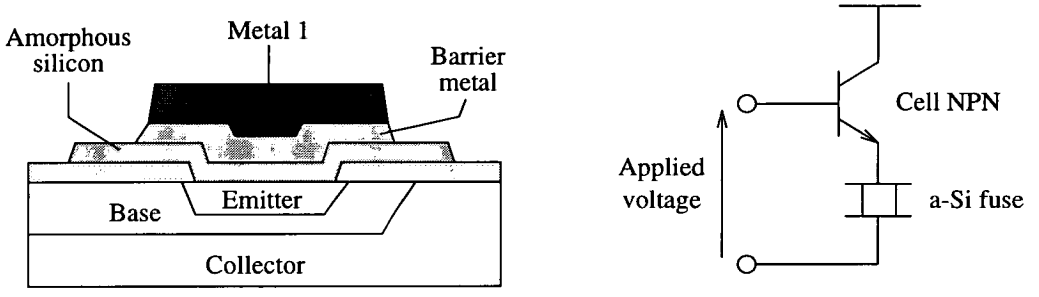
The market for fuse based PROMS is much smaller than that for EPROM and this is not entirely due to the fuse based PROM's lack of reprogrammability. There are in fact One-Time Programmable (OTP) EPROMs on the market which have no UV-window, so making packaging considerably cheaper.

The disadvantage of fuse based technologies is that the basic cells are much larger than the equivalent floating gate devices; this means that the price and available densities are not nearly so competitive. A fuse element is larger for two main reasons: firstly, it is not merged with the select transistor and secondly, fusing requires much larger currents than a minimum sized MOSFET can supply. The advantage of fuses is that they have much lower resistances than an EPROM transistor, so allowing faster switching.

2.2.2.2. Antifuse technologies

Antifuses have found application in FPGAs because of their small area and low parasitic resistance[27]. They also have the advantage of being "normally off" devices so only a small number - about 2% - need to be programmed for a typical application[38]. There are two main categories of antifuse: amorphous silicon and dielectric based.

An amorphous silicon antifuse is made from a thin layer of undoped material, which can be integrated with an NPN driver transistor, as shown in figure 2.10(a). The barrier metal (Ti:W) is used to prevent aluminium spiking[39].



(a) Cross section (b) Symbol

Figure 2.10 - Amorphous silicon antifuse

The antifuse is programmed from an open state to a conducting state by applying a bias voltage high enough to cause destructive breakdown. The currents required during programming are much lower than those needed for fusible link technologies, so smaller driver transistors can be used.

As with the polysilicon fuse the device goes through an intermediate state prior to entering the highly conducting regime. Figure 2.11 illustrates the three different states of a 500Å a-Si antifuse, from initial high resistance, through secondary leakage, to the highly conducting post fuse state.

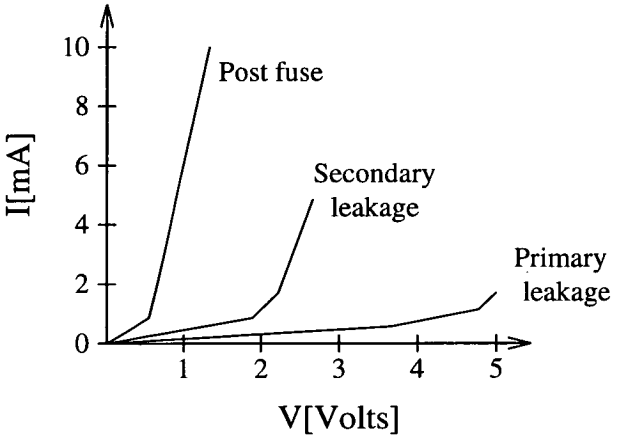


Figure 2.11 - Programming characteristic for a 500Å a-Si antifuse

The use of amorphous silicon antifuses has been hampered by two difficulties: firstly, application of a reverse current can return a programmed antifuse to a nonconductive state and secondly, even unprogrammed devices pass a small but significant leakage current.

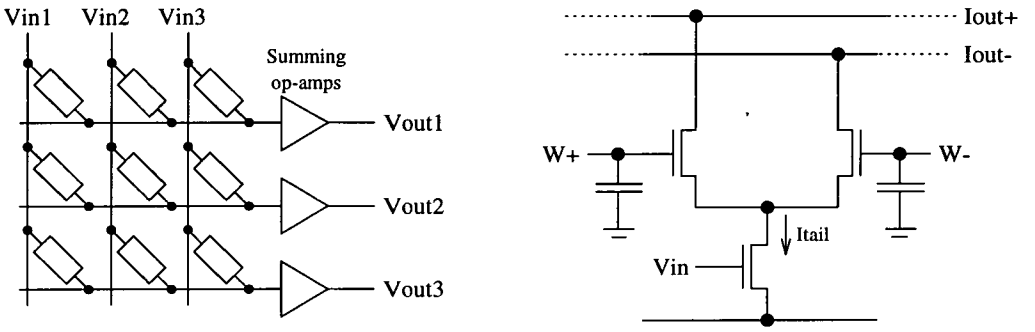
Both these problems can be overcome by using dielectric antifuses. These devices consist of a layer of dielectric sandwiched between N+ diffusion and polysilicon. Upon application of a sufficiently high voltage the dielectric breaks down and the device becomes

highly conductive. An FPGA chip, based on this technology, containing 750,000 anti-fuses has been constructed[38].

2.3. Non-volatile synaptic weight storage

This section considers various technologies used for synaptic weight storage. Memory technology aside, the actual synapse circuit itself usually belongs to one of two general types.

The first common circuit topology is simply an array of conducting elements, as shown in figure 2.12(a). In this arrangement each resistor acts as a synaptic weight: if the resistor has a high value then the coupling from the input to the neuron is small. If the resistor has a low value then the coupling from the input to the neuron is large.



(a) Network based on a resistive array (b) Synapse based on a differential stage

Figure 2.12 - Two common synapse circuit topologies

The second common topology is based on an array of synapses constructed from differential stages, as shown in figure 2.12(b). The transfer function of the differential stage means that the tail current is effectively multiplied by the stage's differential input voltage. In the cell shown in figure 2.12(b) the synaptic weight is stored dynamically on capacitors[40].

As the synapse array forms the bulk of any hardware neural network the choice of memory technology is a crucial one. Designs in which the synaptic weight is defined by the final mask layer, equivalent to ROMs, are extremely compact but are not reprogrammable. At the other extreme there are EEPROM based designs which can be reprogrammed but occupy a much larger area. The following review commences with hard-wired designs and works through technologies equivalent to OTP and EPROM, before considering fully reprogrammable devices, such as EEPROM.

However, as with the digital review, the first technology to be considered is battery backed SRAM.

2.3.1. Battery backed SRAM

The Kakadu ANN chip, designed by Jabri[3], has been designed to suit a very specific application; it implements a two layer network with 10 inputs, 6 hidden nodes and 4 outputs (10,6,4). The chip is intended for use as the classifier in an implantable cardioverter-defibrillator and has therefore to meet a very strict power specification; the chip has a power dissipation of 20 nW.

On the Kakadu chip the synaptic weight is stored using SRAM cells connected to a multiplying digital to analogue converter (MDAC), as shown in figure 2.13. The synapse array thus appears to the controlling digital circuitry as a large write only register.

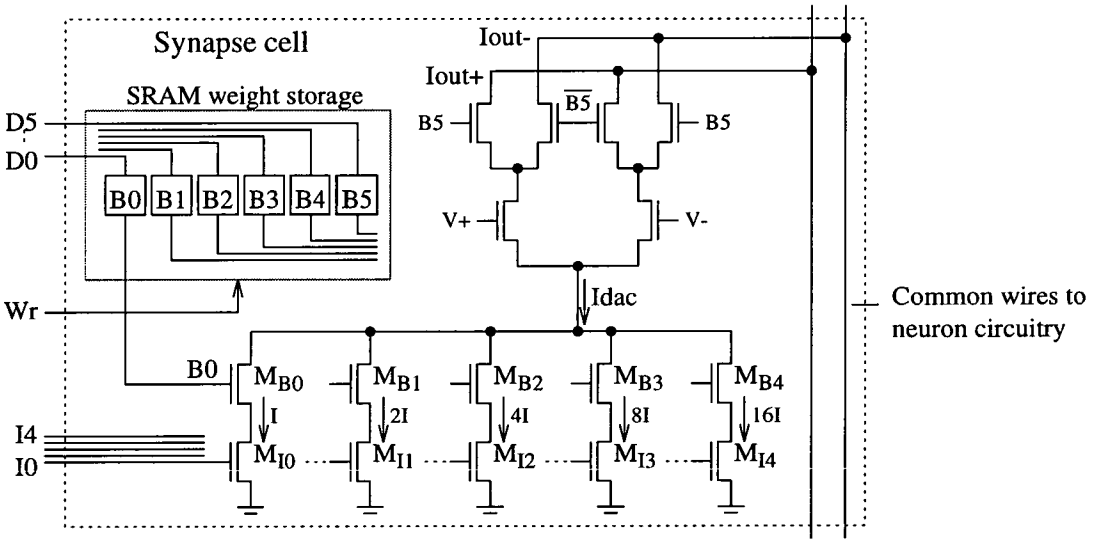


Figure 2.13 - Synapse based on a SRAM storage

The characteristic of the differential stage means that the current generated by the MDAC is effectively multiplied by the differential input voltage,V+ minus V-. The resulting differential output current is connected to the neuron's inputs. In order to obtain full four quadrant multiplication the sign of the synapse weight (B5) controls additional switch logic; if the weight is negative the differential output currents are reversed. The complete synapse occupies an area of 106μm x 113μm.

Although it is not a non-volatile chip the Kakadu design provides an excellent example of a niche application that has been filled by a hardware ANN.

2.3.2. Hardwired synapse arrays

The ANN chips with the highest synaptic density are those based on fixed weight arrays. In these chips the value to be stored is defined by the final mask layer, as in a digital ROM. The majority of the chips discussed in this section implement so called associative memory networks in which an input vector is compared with those in memory and the closest match determined. Associative memories are amenable to implementation with

fixed weight networks for two reasons: firstly, the weights are binary (0 or 1) and secondly, the vectors to be stored in memory are known beforehand.

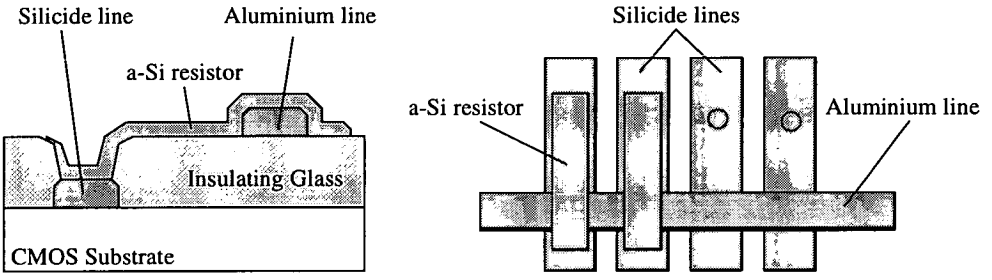
A number of research groups are using resistor based technologies to provide the fixed weight array; all these groups use amorphous silicon either because of its high resistivity or its photoconductive properties.

2.3.2.1. Fixed weight arrays using high value resistors

For a chip containing a massively parallel array of conducting elements to have a reasonable power dissipation, the current per synapse must be extremely low. High value resistors are therefore required. Amorphous silicon is widely used to provide such high value resistors because of its ease of manufacture and extremely low conductivity.

The goal is to make the resistors as small as possible, effectively the crossing point of the row and column address lines. In the chip designed by Hubbard [41] - a 22 x 22 resistor array - the tungsten wires used for row and column addressing have a track width of 2μm and the resulting a-Si resistors have values in the 100s kΩ.

In the chip developed by Graf[42] the a-Si resistor is placed between aluminium and silicide address lines, as shown in figure 2.14(a).



(a) Cross section through resistor (b) Plan view of resistors

Figure 2.14 - Graf style resistor

The binary weights (0 or 1) are set by the presence of an a-Si resistor on the final mask layout, as shown in figure 2.14(b). This chip has 130,000 resistor sites, each occupying an area of only 0.25μm x 0.25μm. The array is connected to 256 invertors, constructed using CMOS technology, that act as summing neurons.

An alternative to using a CMOS backplane is to construct both the resistors and the neurons using thin film technology; using this approach the chip area is not restricted by the limits imposed by crystalline technology. In the chip designed by Busta[43], phosphorus doped a-Si:H was used to construct the synapse weight resistors and thin-film transistors (TFTs) the invertor/op-amps. The network was constructed on a glass substrate using a fabrication process that only requires four mask stages.

2.3.2.2. Fixed weight arrays with optical inputs

The photoconductive properties of a-Si mean that resistors fabricated from this material can be used to provide an optical input to hardware ANNs. In the network proposed by Binns[44] the a-Si synapse resistor is defined by the overlap of metal and transparent ITO (Indium Tin Oxide) tracks, as shown in figure 2.15.

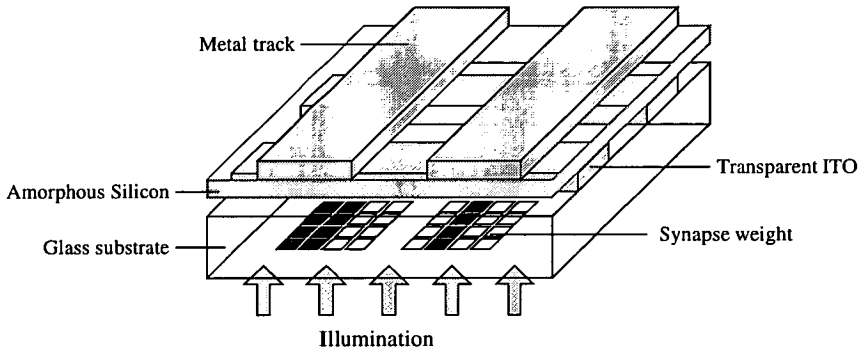


Figure 2.15 - a-Si resistor array

The synaptic weight, calculated using a software simulation of the complete network, is defined by a 4x4 grid of pixels (16 levels) drawn on the substrate surface below the synapse site. The neurons are implemented using operational amplifiers made from poly-Si based TFTs.

Kornfeld[45] has also constructed a network based on an amorphous silicon photosensor array this time with 14,400 synapses. Photographic film is used to define the weight set.

2.3.2.3. Fixed weight arrays using capacitors

Another approach to the implementation of fixed weight networks is to use capacitors, the area of which determines the synaptic coupling. As the resulting matrix is devoid of active devices it offers very high synaptic space-power efficiency. The network designed by Cilingiroglu[46] uses a three phase clocking system to transfer charge from the voltage inputs, through the synaptic array, to the output nodes. The capacitor in the synapse cell is either connected to the $\phi 1$ or $\phi 2$ line depending on whether it represents a positive or negative weight, as shown in figure 2.16.

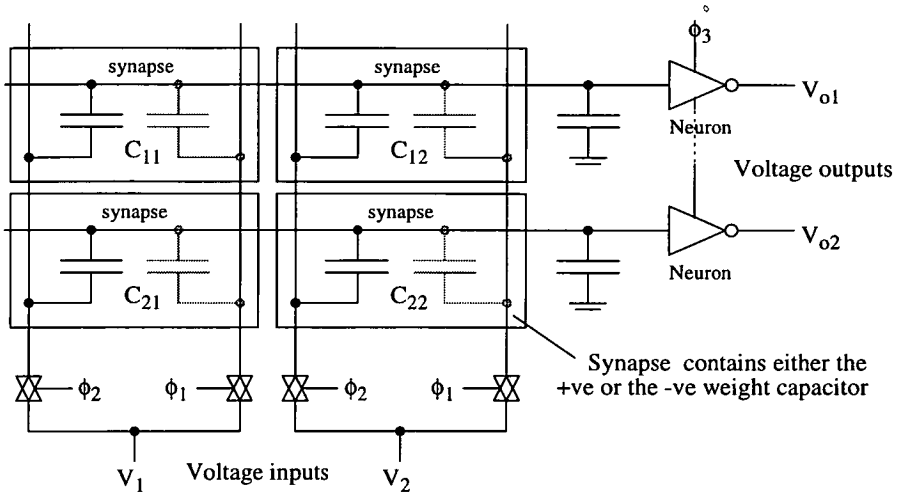
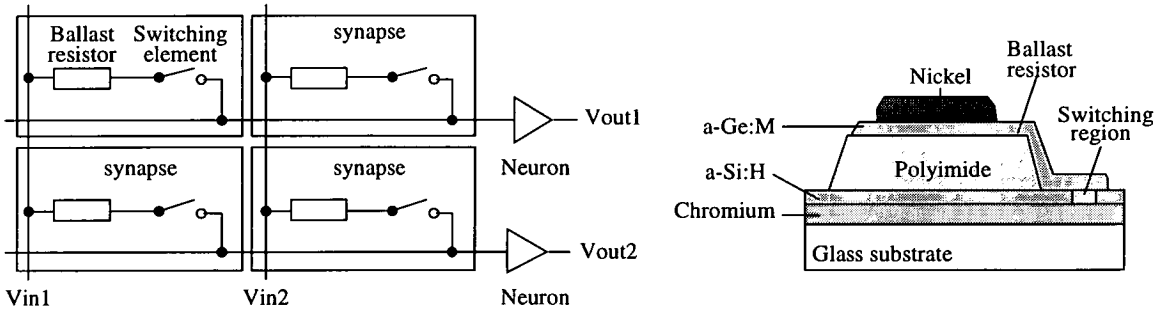


Figure 2.16 - Fixed capacitor synapse array

Although the size of the synapse cell, 16.5μm x 10μm in 3μm technology, is larger than that used in resistor based designs, this approach has the advantage of not requiring any specialised processing. The network described in Cilingiroglu's original paper was restricted to binary weights but the synapses could easily have analogue values.

2.3.3. PROM - Write Once synapse array

Having considered fixed weight networks, the equivalent of digital ROM chips, the next network to be considered is one time programmable, the equivalent of digital PROM devices. In the network described by Thakoor [47] each synapse cell contains a switchable element and a ballast resistor, as shown in figure 2.17(a).



(a) One time programmable binary network (b) Cross section through synapse

Figure 2.17 - Network based on a-Si switch

The ballast resistor provides the connection strength and "limits" the energy delivered during the switching process. As this network is intended for large associative arrays the ballast resistor needs to provide a very weak connection, greater than 10^6 ohms for a 1000 x 1000 matrix. Various ballast resistor materials based on amorphous semiconductors and cermet systems were tried, the major constraints being a need for a small feature size and thermal stability[48]. In the cell shown in figure 2.17(b) the ballast resistor is made from amorphous germanium (a-Ge:Cu, a-Ge:Al) and the switching element from

r.f. sputtered hydrogenated amorphous silicon (a-Si:H).

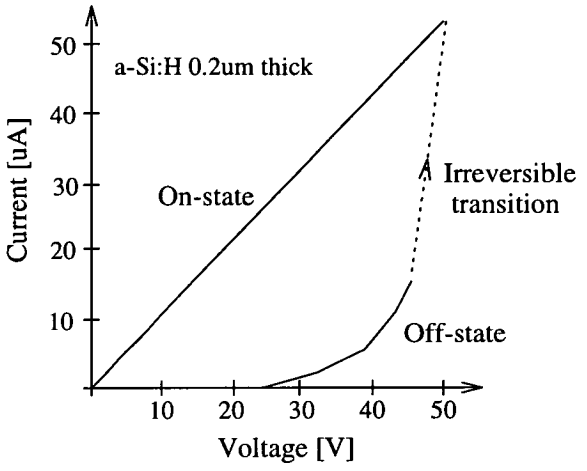


Figure 2.18 - Irreversible switching in a-Si:H thin film. After Thakoor[48]

The memory switching in the a-Si:H is current induced and requires very low energy. The cause of switching is believed to be the formation of a low resistivity crystalline filament. As figure 2.18 shows, the switching voltages needed to induce the state transition are very high, 20 V to 50 V.

2.3.4. EPROM equivalent - The UV-memory

The UV-light used to erase EPROMS can also be used to inject electrons onto a floating gate. During the UV-illumination there is effectively a small leakage conductance in parallel with the floating gate capacitor. This conductance can be used to change the charge incrementally on the floating gate. Mead used such a device[49] to adapt the offsets in his artificial retina chip. An advantage of this technology over EEPROM is that the polysilicon layers available in a standard CMOS process can be used for the floating gate. The thicker oxide also means that the charge will remain for years once the UV-illumination is removed[6].

One of the problems with this approach is that UV-induced conductances will arise between all layers separated by SiO_2. This means that all non UV-structures must be shielded against UV-light[50]. This is accomplished using the metal 2 layer as a shield, with openings above the UV-structures, as shown in figure 2.19.

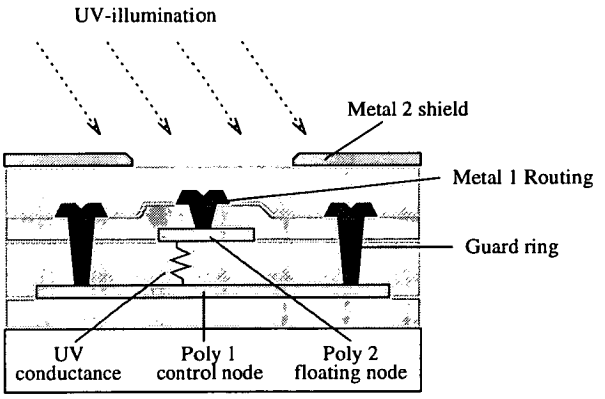


Figure 2.19 - UV memory structure

At Caltech, Cauwenberghs[51] has designed a compact (30μm x 30μm) two transistor synapse based on UV-programming. The synapse, shown in figure 2.20, contains a memory transistor, used to store the synaptic weight, and an access transistor, used to apply the programming voltages.

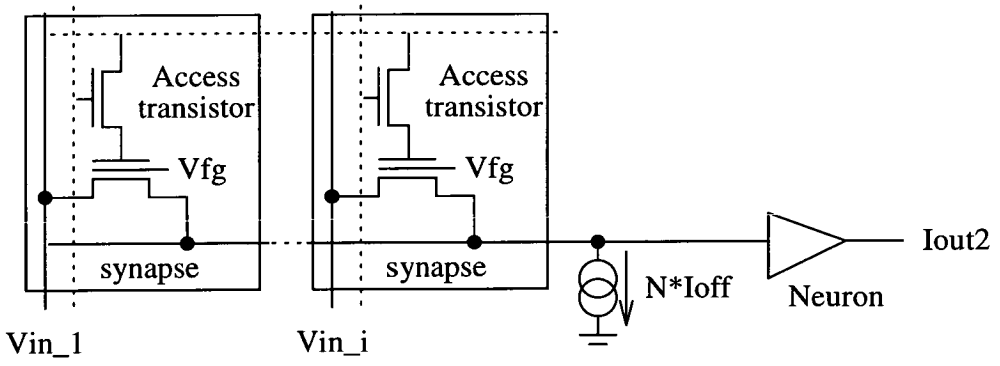


Figure 2.20 - Network based on UV-analogue memory

By ensuring that the input voltages are small, -400 mV to 400 mV, the floating gate transistor remains in its linear region where the output current is proportional to the gate voltage. The output current is therefore a product of the input voltage and the weight stored on the floating gate. To obtain four-quadrant multiplication a constant bias current is subtracted from the output of each synapse; this is implemented as a single unit at the foot of the column as shown in figure 2.20.

Tawel and Thakoor have also constructed UV-programmable synapses based on a more conventional 4-quadrant multiplier[52].

By using these UV-structures it is possible to construct synapse cells with long term storage using a standard CMOS process. The disadvantage, as with EPROM, is the need for an external UV-light source.

2.3.5. Fully programmable (reversible) technologies

Having now considered the neural equivalents of ROM and EPROM the next set of technologies to consider are those that are fully reprogrammable. ANNs with reprogrammable analogue weights are almost all based on proven digital memory technologies. However, there are some programmable resistor technologies that are unique to the field of artificial neural networks. The different technologies that will be considered are:

- FLOTOX EEPROM
- MNOS and SONOS
- Ferroelectric
- Programmable resistor technologies

2.3.5.1. EEPROM and Floating Gate

In recent years there have been a number of papers in which EEPROM devices have been used for non-volatile, analogue storage. As these are not all explicitly neural they have been included in Appendix A, along with articles on floating gate cells constructed using a standard CMOS process.

The best known hardware neural net chip is probably the ETANN (Electronically Trainable ANN) chip developed by Intel[53]. The ETANN chip contains 10240 synapses each of which uses two EEPROM transistors for analogue, non-volatile storage. The synapse, which has an area on 41.6μm x 48.3μm, is based on a 4-quadrant multiplier, the tail current of which is determined by the floating gate devices, as shown in figure 2.21.

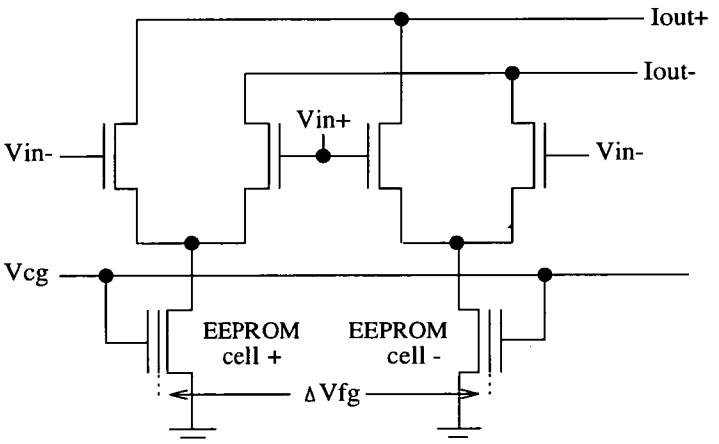
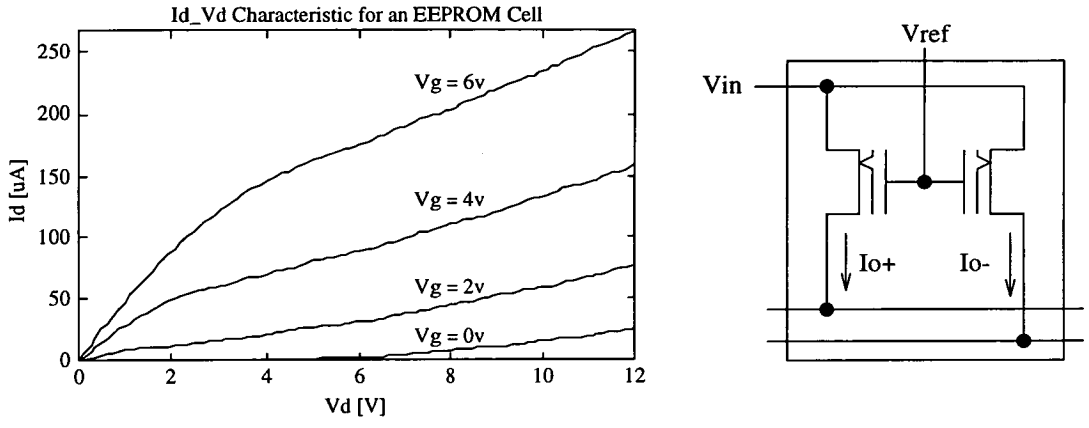


Figure 2.21 - ETANN synapse

To study the weight retention performance of the synapse cells they were baked at 250°C: 3200 minutes at 250°C is equivalent to 15 years at 125°C. From the results of these experiments the weight accuracy was estimated to be 6-7%, equivalent to 4-bits[54]. The programming pulses are 10 μs to 1 ms in duration, with the height of the pulse calculated from the target weight and the charge already stored.

Kramer[55] has designed a floating gate synapse that only requires two transistors per cell. It takes advantage of the EEPROM's unusual Vds/Ids characteristic: the high tunnelling capacitance between the gate and the drain makes the Ids current in the saturation region strongly dependent on Vds, as illustrated in figure 2.22(a).



(a) EEPROM Vds/Ids characteristic (b) 2 transistor synapse

Figure 2.22 - Compact two transistor EEPROM synapse cell

By using this characteristic to perform the multiply operation it is possible to get a two transistor synapse cell, the area of which is only 200 μm^2, less than one tenth that of the ETANN synapse. The synaptic weight is stored as the transistor's threshold voltage. The voltage input is then used to supply the Vds voltage.

A single transistor would only provide one quadrant multiplication. Two EEPROMS can be used to provide 2-quadrant multiplication using a common-input, differential output current scheme as shown in figure 2.22(b).

Shimbukuro[56, 57] uses a four transistor synapse based on hot-electron programming to achieve programming with lower voltages, in the range 12 V to 20 V.

2.3.5.2. Silicon Nitride - MNOS and SONOS

After demonstrating that it was possible to store analogue values using MNOS capacitors[58] Withers and Sage then used MNOS devices for the non-volatile storage of synaptic weights[59]. A 13 x 13 network that used charge-coupled devices (CCDs) for computation and MNOS for storage was built[60]. In this approach the synaptic weight is stored as a charge on a capacitor and it is packets of charge that are summed rather than currents. The main practical problem with using MNOS memory cells is that voltages of +35 V and -35 V are required for programming.

The modern variant of MNOS is SONOS; it has an additional oxide blocking layer that allows much lower programming voltages compatible with CMOS circuitry. Figure 2.23 shows the programming characteristics of a SONOS based synapse developed by White and Chen[31].

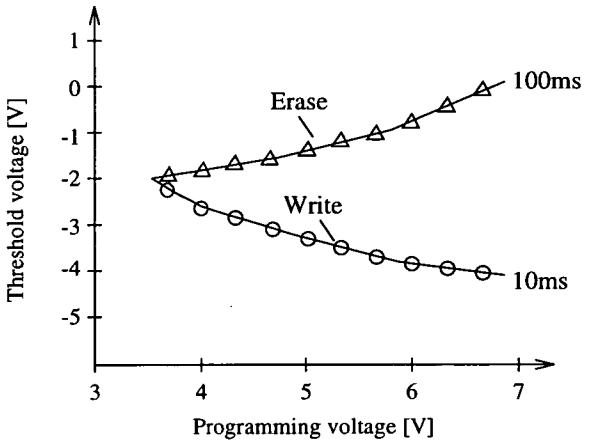


Figure 2.23 - SONOS synapse programming characteristics. After White[31]

As figure 2.23 shows, the SONOS synapse can be programmed with voltages in the range 3 V to 7 V, considerably lower than those required for EEPROM cells. The SONOS memory is extremely compact, occupying an area of 5μm x 5μm, and has estimated weight decay of 20% over a projected 10 year period.

2.3.5.3. Ferroelectric capacitors

An analogue synapse cell has been constructed using the ferroelectric thin film capacitors discussed in the section on digital storage[61]. However, in this case it is a variable charge that is stored on the capacitor, allowing analogue storage. This stored charge is used to generate the input voltage to a multiplier cell, as shown in figure 2.24.

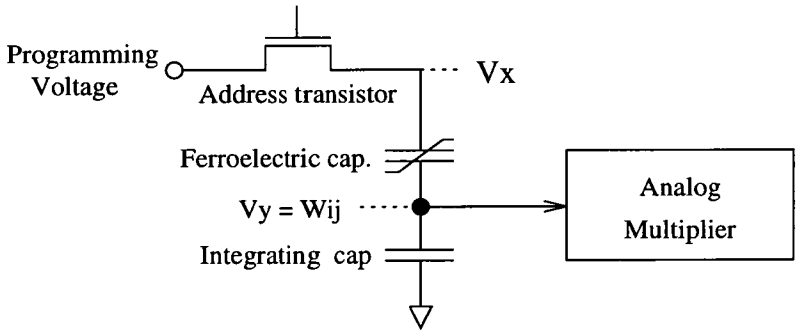


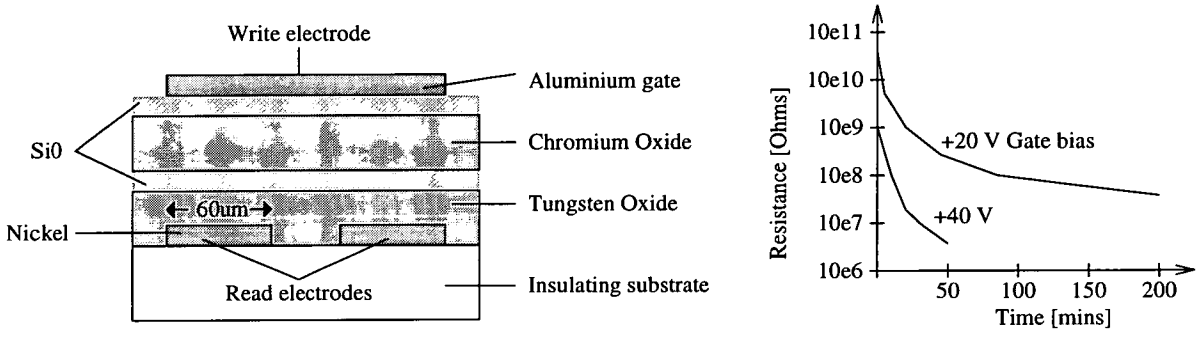
Figure 2.24 - Ferroelectric synapse cell

Unlike the digital ferroelectric cell the ferroelectric synapse cell uses a non-destructive readout circuit; this removes the need for refresh circuitry. At this time ferroelectrics are still an immature technology and today's devices suffer from gradual weight decay over time.

2.3.5.4. Programmable resistor technologies

One group of non-volatile, reprogrammable technologies unique to the field of neural networks are the modifiable resistor (memistor) devices. In this section three different memistor technologies will be discussed. The first of these is based on thin films of tungsten oxide.

• Tungsten Oxide: Ramesham[62] has constructed a solid-state resistor based on thermally evaporated, tungsten oxide thin films[60]. The resistance of the tungsten oxide can be reversibly modified by injecting cations (H^+, Na^+) from a solid state electrolyte using a third control gate electrode. In the device shown in figure 2.25(a) chromium oxide is the electrolyte that acts as the ion source. Silicon oxide is used as a blocking layer between the tungsten oxide and the ion source.



(a) Cross-section (b) Programming characteristics

Figure 2.25 - Tungsten Oxide electrochemically reprogrammable device

Figure 2.25(b) is the device programming characteristic. As this shows, the device resistance can be switched over four orders of magnitude. However, the programming pulses are 20 V to 40 V high and have duration in the order of minutes[63].

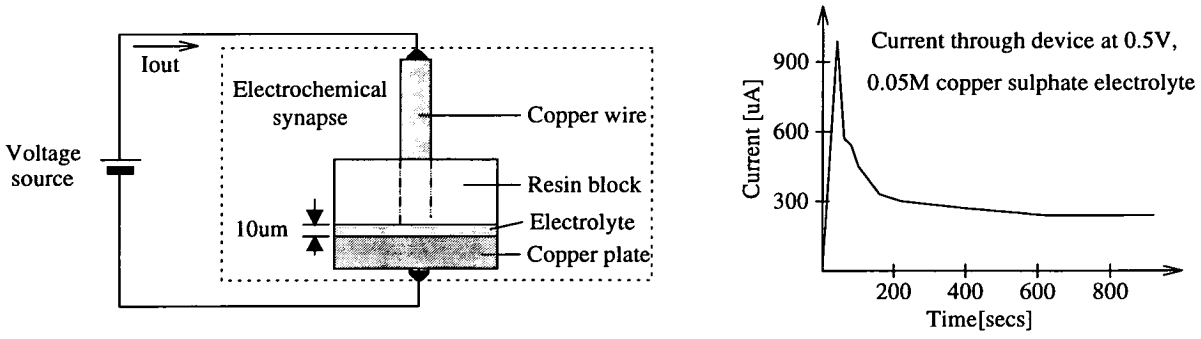
• Bismuth Oxide: Spencer [64] has investigated the potential of bismuth oxide as a programmable resistor technology. The compounds investigated include Bi_2O_3 and $Bi_{12}GeO_{20}$.

Switching and programming has been observed in samples of these compounds with voltages as low as 50 mV and with currents in the nanoampere range. The resistors have values of 10^6 to 10^9 ohms in the off-state and 10^4 to 10^5 ohms in the on-state. Spencer suggests that the memory action is based on the creation of microscopic conducting paths in the material; reverse polarity pulses interrupt these conducting paths and return the device to its more insulating state.

Since the original paper there have been no further reports on the progress made with this programmable resistor technology.

• The electrochemical synapse: Triffet [65] has constructed an electrochemical synapse in the hope of reproducing the temporal aspects of neural signals. The synapse consists of

copper electrodes and a copper sulphate electrolyte, as shown in figure 2.26(a).



(a) Test setup (b) Typical results

Figure 2.26 - Electrochemical synapse cell

When a voltage is applied to the cell the current pulse that is produced, shown in figure 2.26(b), has a shape similar to that seen in neural signals. The next stage in the development of this electrochemical synapse is intended to be an integrated circuit version of the test cell shown in figure 2.26(a).

2.3.6. Beyond EEPROM - Self programmable arrays

While the review of digital memory devices concluded with EEPROM - electronically programmable and erasable chips - a review of synaptic weight storage must also include memory devices suitable for On-Chip Learning (OCL). In an OCL system the chip is provided with a set of inputs and desired outputs and is then left to evolve a weight set that will perform the mapping function between them.

FLOTOX EEPROM is a good candidate for the synapse element in an OCL chip[66] because the stored charge, and hence the synaptic weight, can be incremented gradually, a requirement of some on-chip learning algorithms[67].

Montalvo[68] has proposed a temperature compensated OCL chip that uses dynamic storage during the learning phase and EEPROMs for long term storage.

2.4. The a-Si:H analogue memory - Introduction

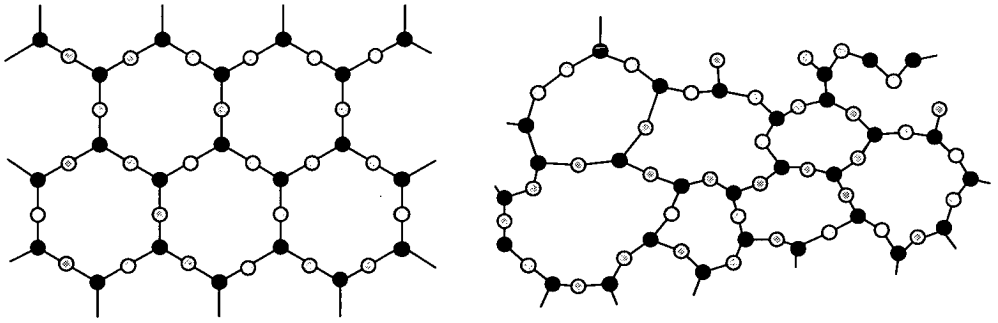
Having considered various digital and analogue non-volatile technologies this final section contains a brief description of the a-Si:H analogue non-volatile memory, developed jointly by Dundee and Edinburgh Universities. The discussion is divided into four main sub-sections:

- Amorphous materials
- Switching in amorphous materials
- The a-Si:H analogue memory

- The BT a-Si:H XOR demonstrator

2.4.1. Amorphous materials

The principle difference between crystalline and amorphous materials is their structure: the atoms that constitute a crystalline material are aligned in a regular crystal lattice, while in an amorphous material there is no such well defined long range order. Figure 2.27 illustrates the difference in structure between a crystalline and an amorphous material[69].



(a) Crystalline structure (b) Amorphous structure

Figure 2.27 - Crystalline and amorphous materials

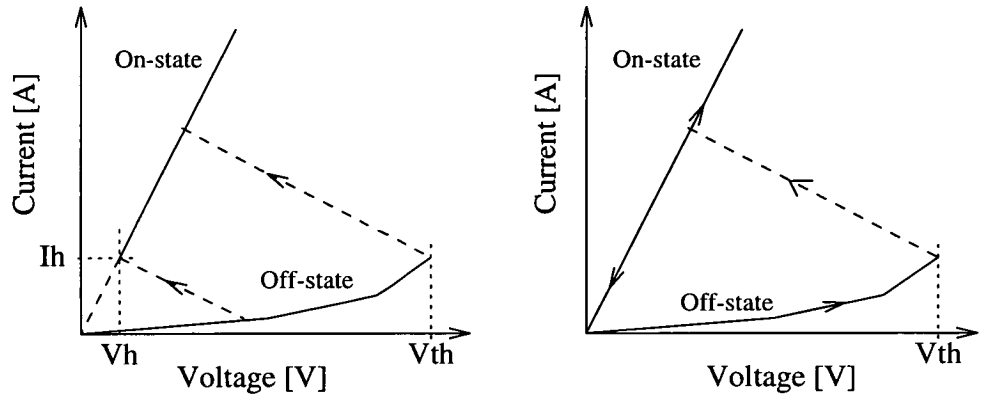
There are two main classes of amorphous semiconductors: chalcogenide glasses and amorphous solids. The chalcogenide glasses, so called because they contain a high percentage of one of the "chalcogen" elements (sulphur, selenium and tellurium), exhibit many interesting properties, including reversible phase transition and photoelectric effects. The amorphous solids, amorphous silicon (a-Si) and amorphous germanium (a-Ge), have properties similar to their crystalline counterparts; they can be doped with impurities to form solid state pn structures (diodes and transistors).

The advantage of amorphous materials when compared with crystalline ones is their relative ease of manufacture. Crystalline materials require high temperatures and slow growth in order to produce a high quality crystal lattice. By comparison, thin films of amorphous semiconductors can be grown as a coating using transition from the vapour phase. This allows large areas to be covered with amorphous material and has resulted in a-Si transistors being used as the driver transistors in Liquid Crystal Displays (LCDs). In addition, the photoconductive properties of amorphous silicon have resulted in its use in applications such as solar cells and image sensors.

2.4.2. Switching in amorphous materials

Switching in amorphous devices can be divided into two classes[70]: threshold switching and memory switching. A device that exhibits threshold switching changes from its off-state to its on-state if the applied voltage exceeds a threshold value. If the on-state then

falls below a holdpoint (Ih,Vh) the device reverts to its off-state, as illustrated in figure 2.28(a).



(a) Threshold switching (b) Memory switching

Figure 2.28 - Threshold and memory Switching

Threshold switching is nonpermanent or "volatile" as it always reverts to the off-state in the absence of an applied bias. In a device that exhibits memory switching both the on-state and the off-state characteristics extrapolate through the I-V origin. The on-state is thus retained once the bias is removed, giving a permanent or non-volatile memory action, as illustrated in figure 2.28(b).

In the 1970s many examples of switching in amorphous materials were reported, most importantly in chalcogenide glasses, where both threshold and memory switching were observed. Threshold switching was also observed in vacuum evaporated films of a-Si; this was interpreted as being electrothermal, involving a conducting filament similar to that seen in chalcogenide glasses.

In 1982 three groups published work on switching in hydrogenated a-Si films. Hydrogenated amorphous silicon (a-Si:H) is obtained by the glow discharge of Silane. Figure 2.29 shows the switching characteristic of the p^+ni device developed by Edinburgh and Dundee Universities.

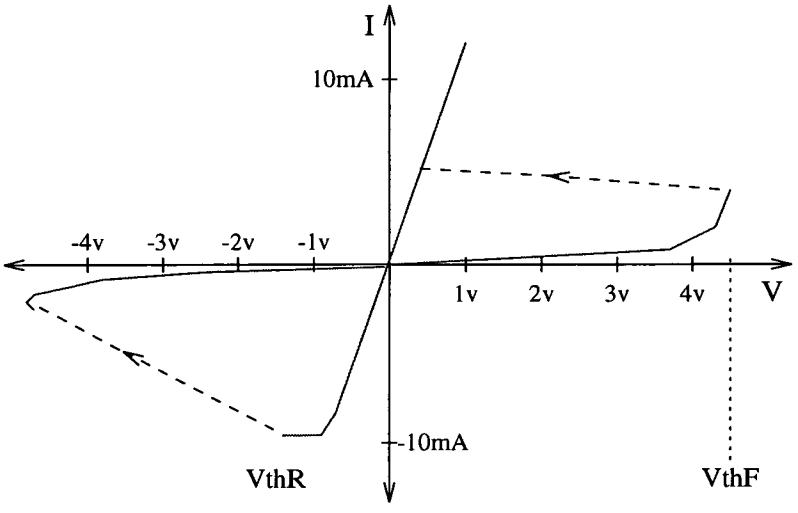


Figure 2.29 - Static I-V characteristic of a formed a-Si pni device. After Hajto[69]

As figure 2.29 illustrates, the device's digital memory switching is reversible: by applying a negative bias the device can be switched from the conducting on-state back to the off-state. This switching behaviour differs from that seen in chalcogenide glasses in two regards: firstly, it is polarity dependent and secondly, much less energy is required, 1 μJ compared with 1 mJ.

2.4.3. The a-Si:H analogue memory device

During research on the a-Si:H digital memory many different cell structures were experimented with. The two that were the most consistently reliable, over 10^6 switching cycles, were M-p$^+$ni-M' structures, discussed in the last section, and M-p$^+$-M' structures, M and M' being the metals used to construct the electrodes. The operation of both these structures was found to be dependent on the metal used for the top electrode.

When chromium was used as the top electrode the device displayed a digital memory action i.e. it switched between a stable on-state and a stable off-state. However, when other metals, such as vanadium, were used the resulting device displayed a non-volatile, analogue memory behaviour i.e. the device could be switched into a variety of stable intermediate resistance states.

The new resistance state was determined by the height of the programming pulse applied to the device; the range of programming voltages that could be applied was referred to as the programming window. The width of this programming window ΔV_p depended on the metal used for the top electrode: for an analogue device the programming window was large† (vanadium, ΔV_p = 1.5 V), while for a digital device it was very narrow (chromium, ΔV_p = 0.2 V). Table 2.1 summarises the switching associated with different

† Figure 2.31(a) shows the switching characteristic of an analogue device illustrating the wide programming window.

top metals.

Metal	ΔV_p[V]	Switching characteristics
Al	0.1	Digital, non-volatile
Cr	0.2	Digital, non-volatile
V	1.8	Analogue, non-volatile
Ni	2.0	Analogue, non-volatile
Co	2.0	Analogue, non-volatile
Mo	2.0	Analogue, volatile

Table 2.1 - Effect of top metal on switching behaviour. After Hajto[71]

The research at Edinburgh and Dundee Universities is currently based on a structure consisting of a vanadium top electrode, a 1000Å layer of p^+ a-Si:H, and a chromium bottom electrode, as shown in figure 2.30. The active pore of the device is defined by a layer of baked photoresist.

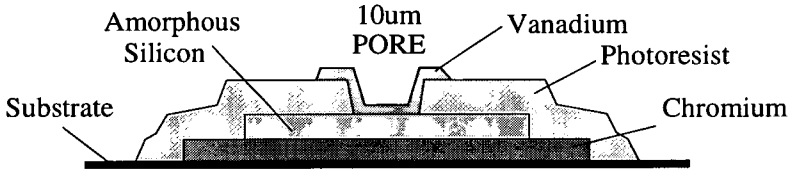


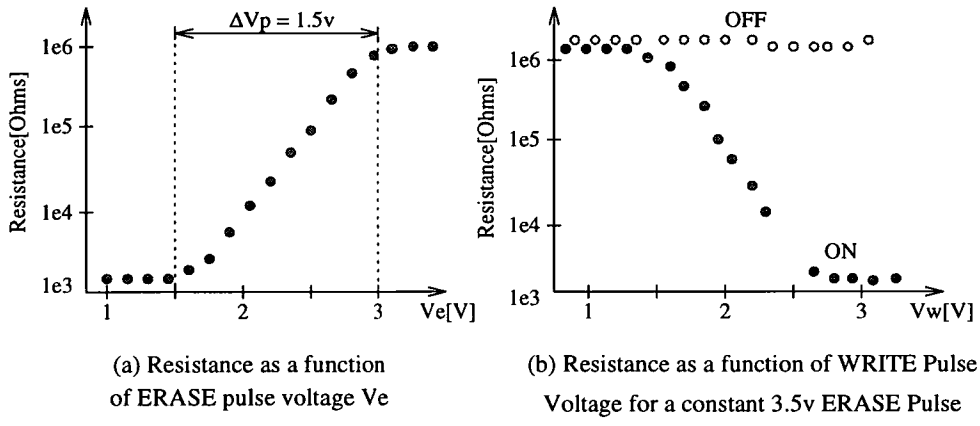
Figure 2.30 - Cross section of a-Si:H memory device on glass. After Hajto[72]

After fabrication the a-Si:H device has a very high resistance (several GΩ), owing to the metal-semiconductor Schottky barriers at the contacts. To program the device into a lower resistance state the following steps must be carried out:

- Forming: This is a once only process. A series of positive† 300 ns pulses increasing in amplitude from 5 V to 14 V, is applied across the device electrodes. This creates a vertical conducting channel which can be programmed to a value in the range 1 kΩ to 1 MΩ.
- Write: To decrease the device resistance , negative, "write", pulses are applied.
- Erase: To increase the device resistance , positive, "erase", pulses are applied.
- Read: The device resistance can be "read" using a voltage of less than 0.5 V without causing reprogramming.

The programming pulses (write or erase), which range between 2 V and 5 V, are typically 120 ns in duration. Figure 2.31 illustrates the effect of a series of write and erase pulses on the device resistance.

† A positive pulse is defined as one where the top (vanadium) electrode is at a higher potential then the bottom (chromium) electrode.



(a) Resistance as a function of ERASE pulse voltage Ve

(b) Resistance as a function of WRITE Pulse Voltage for a constant 3.5v ERASE Pulse

Figure 2.31 - Resistance as a function of pulse height. After Rose[73]

In figure 2.31(a) the device resistance is seen to change between 1 kΩ and 1 MΩ depending on the height of the erase pulse; the programming window is 1.5 V.

In figure 2.31(b) the magnitude of the write pulse is used to set the final device resistance. In this example, the device is "reset" to its off state (high resistance) prior to each write pulse. Reeder[74] claims that the a-Si:H memory can be programmed with an accuracy of 5% in the range 1 kΩ to 1 MΩ, equivalent to a digital 4-bits.

Although the programming mechanism of the memory devices is not yet understood fully, it is thought that the current in a formed device is carried by a filament which is less than 1μm in diameter. Formation of a filament may be associated with a diffusion of the top metal into the a-Si:H, resulting in a dispersion of metallic atoms in the insulating a-Si:H matrix.

When the I-V characteristics of a-Si:H devices (at room temperature) were investigated it was found that the curves, extrapolated above 1 V, met in a region of about 3 V to 4 V, where switching occurred[71]. This programming or transition regime is illustrated in figure 2.32.

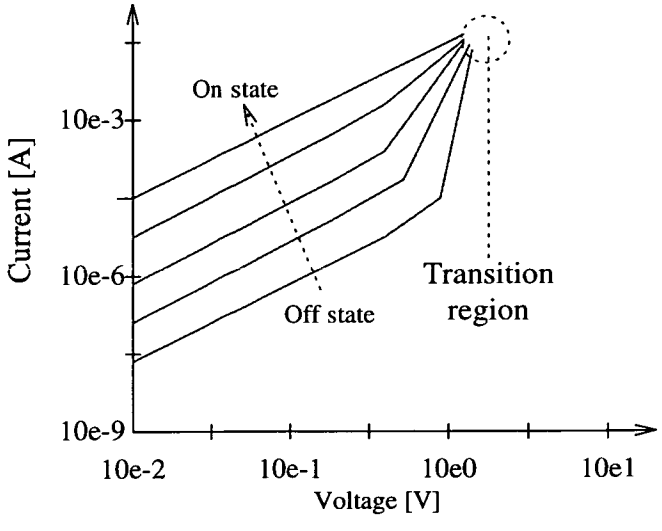


Figure 2.32 - I-V Characteristics of analogue memory states. After Hajto[71]

The nature of the non-volatility in a-Si:H devices was investigated by Rose[13] and was found to be robust against both temperature and radiation induced stress. The resistance of devices was monitored over a period of four years, under zero bias conditions, and found to be stable.

In recent months an Australian group has published results on switching in a-Si:H memory devices that confirms the non-volatile, analogue memory action previously reported by Edinburgh and Dundee Universities[75].

2.4.4. Synaptic weight storage using a-Si:H memory devices

The first practical application of this memory technology was a demonstrator built at the BT Laboratories. A chip containing a 5 x 4 array of a-Si:H devices was used to implement a synaptic weight array that solved the XOR problem[12]. The resistors on the chip were programmed to values determined by a software simulation[76].

Figure 2.33 shows the complete XOR demonstrator which includes the a-Si:H array chip and external op-amp neuron circuitry.

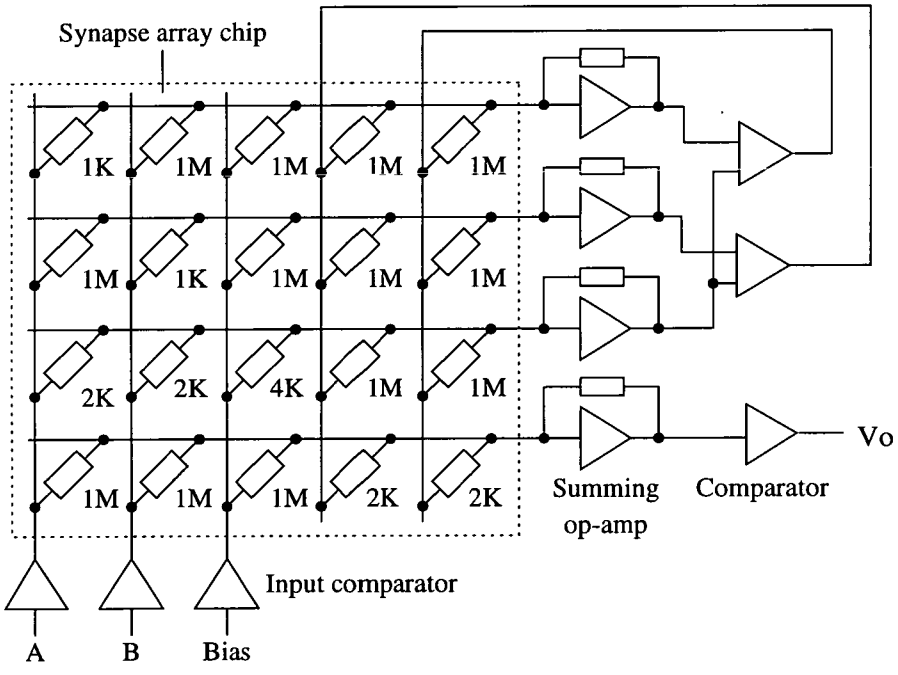


Figure 2.33 - BT XOR demonstrator based on an array of a-Si:H resistors

The operation of the system can be seen more clearly if it is redrawn as a two layer network, highlighting the seven programmed resistors, as in figure 2.34.

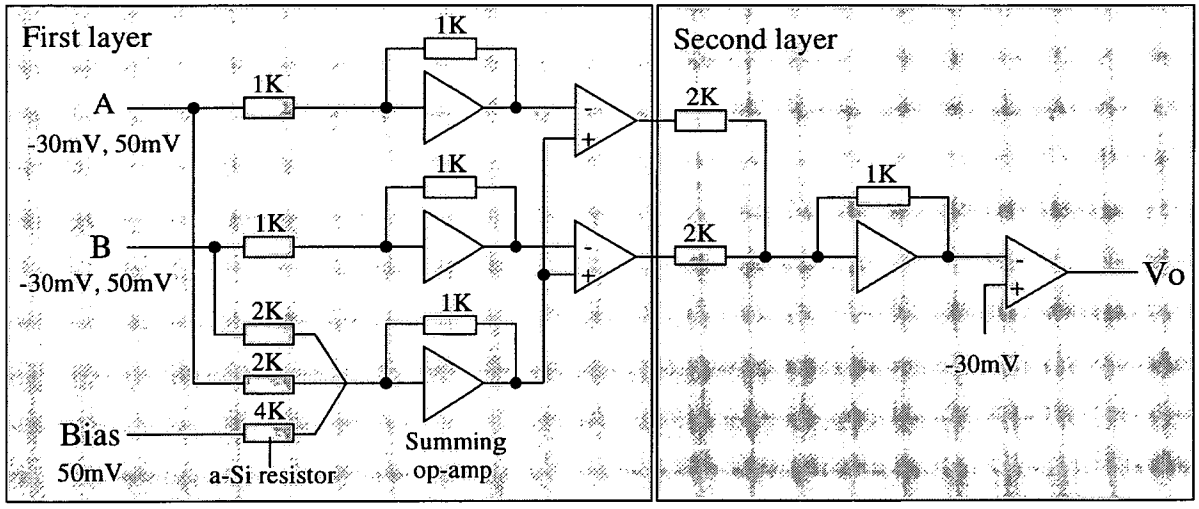


Figure 2.34 - BT XOR demonstrator redrawn as a two-layer network

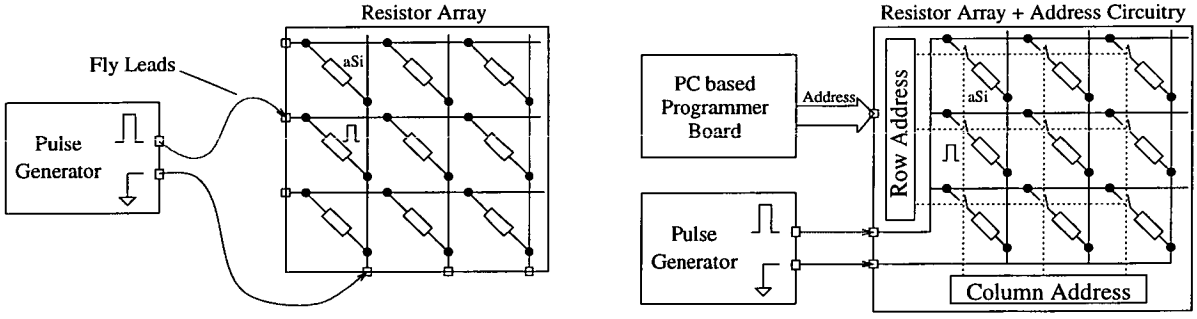
While this network does use a-Si:H for synaptic weight storage there are a number of practical limitations with using a chip containing only an array of resistors. Many of these problems can be addressed by integrating the a-Si:H devices with a CMOS back-plane containing address and synapse circuitry, as detailed in chapters 3-5 of this thesis. In the final chapter elements from both the review material in this chapter and the experimental results from subsequent chapters are brought together in a discussion on possible future directions for non-volatile synaptic weight storage.

Chapter 3

ASiTEST1 - Integrating a-Si:H Memory Devices with CMOS

3.1. Introduction

The BT XOR demonstrator, discussed in chapter 2, was the first chip to use a-Si:H memory devices as a means of storing synaptic weights. While it demonstrated both the analogue behaviour and non-volatility of the a-Si:H memory device, the use of chips that contain only a resistive array is limited by a number of practical considerations. Principal amongst these is the need for external addressing circuitry: for the XOR demonstrator the a-Si:H device to be programmed had to be selected manually, by connecting fly-leads across the appropriate pins on the chip, as shown in figure 3.1(a).



(a) External/Manual Addressing (b) Digitally Controlled Addressing

Figure 3.1 - Addressing on a-Si:H test chips

Obviously, if chips containing large arrays of devices are to be programmed then some form of on-chip addressing circuitry is required. A resistor array chip with digital row/column addressing is illustrated in figure 3.1(b). Now, the pulse generator used for programming is fixed and a digital address is used to select a device from the array.

In the introduction to this thesis it was stated that the primary goal was to replace the capacitor used for dynamic weight storage with an a-Si:H memory device. As the original circuits were fabricated using CMOS technology this was the one chosen to implement the address circuitry on this first testchip. However, the memories themselves are thin-film devices which have the potential of being integrated with other technologies such as thin film transistors (TFTs).

The objective then of the first test chip, ASiTEST1, was to construct a-Si:H memory devices integrated with CMOS address circuitry. The first half of this chapter focuses on

the design of the ASiTEST1 chip, while the second half presents results from the testchip, placing particular emphasis on switching and operating regimes suitable for the memory device. In chapter 4 these results are used as the basis for the design of synapse circuits that use a-Si:H devices for weight storage.

3.2. ASiTEST1 - Design

The two main issues that had to be addressed in the design of the ASiTEST1 chip were:

(i) How to connect the a-Si:H memory device to the underlying CMOS circuitry.

(ii) The design of the address/programmer circuit.

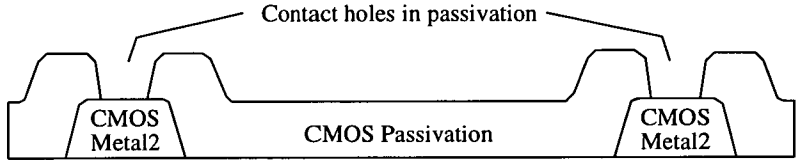
3.2.1. Adding the a-Si:H memory

In a conventional CMOS fabrication cycle the final process step is the addition of a thermal CVD (Chemical Vapour Deposition) silicon nitride, or silicon dioxide, passivation layer across the whole wafer. This is designed to protect the underlying CMOS circuitry during subsequent dicing and bonding operations. Normally, the only openings in the passivation are over the bondpads, to allow the pads to be wire bonded to the chip carrier.

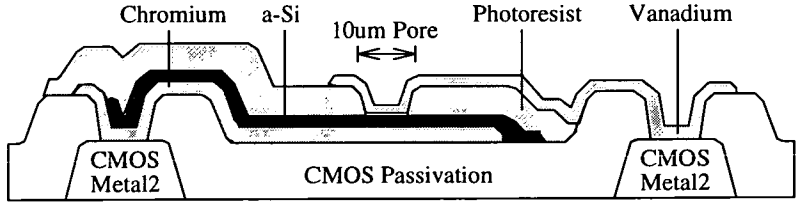
In order to connect an a-Si:H memory device to its CMOS address circuitry additional openings must be included in the passivation. On ASiTEST1 two different types of passivation opening were used:

(i) Contacts in passivation device

In the first approach contact holes were included in the passivation above the metal2 connection nodes, as shown in figure 3.2.



(a) ES2 processing complete



(b) a-Si:H processing complete

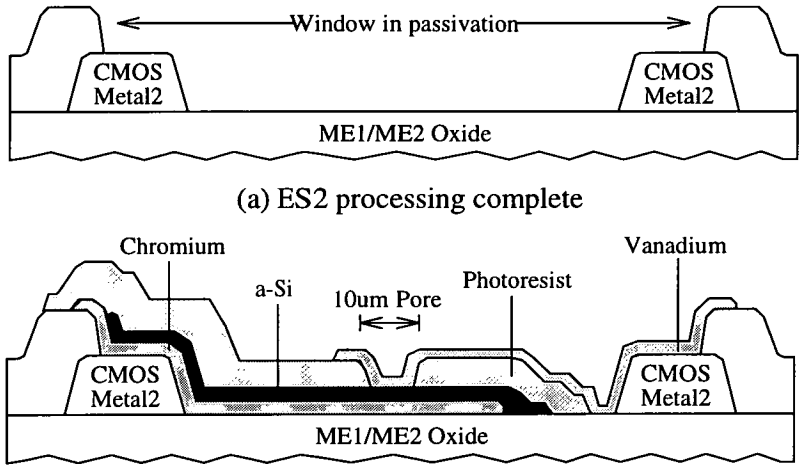
Figure 3.2 - Contact in passivation a-Si:H memory device

One potential problem with this arrangement was the relatively large step between the passivation opening and the metal2 layer. For the process used in the fabrication of the ASiTEST1 chip this step is 1.2μm in height. This must be negotiated by the vanadium, a-

Si:H and chromium tracks which are of the order of 0.1μm thick.

(ii) Window in passivation device

In the second approach the opening in the passivation was along the entire length of the memory device, as shown in figure 3.3.



(a) ES2 processing complete

(b) a-Si:H processing complete

Figure 3.3 - Window in passivation a-Si:H memory device

The design rules for the process used did not allow for openings in the passivation except above metal2. This meant that the etch used to remove the passivation would also remove some of the oxide between the metal2/metal1 layers, causing an additional step that the a-Si:H layers must overcome.

On the ASiTEST1 chip all the test structures are duplicated, one test cell having a "Contact in Passivation" device and the other a "Window in Passivation" device. A more detailed discussion on the procedure for fabricating a-Si:H devices on CMOS wafers, developed in collaboration with Dundee University, can be found in Appendix B.

3.2.2. Addresser circuit design

Having developed two methods for constructing a-Si:H memory devices on the surface of the CMOS wafer, the second major issue to be considered was the design of the addresser circuit. The functions that this circuit had to perform are as follows:

- Apply forming pulses of width 300 ns and amplitude up to 14 V to the addressed device.
- Apply write programming pulse of width 120 ns and amplitude in the range -1 V to -5 V across the addressed device.
- Apply erase programming pulse of width 120 ns and amplitude in the range 1 V to 5 V across the addressed device.

- Ensure that the voltage across a non-addressed device was kept below 0.5 V in order to prevent re-programming.

The simplest method of digitally addressing an array of resistive elements is to place a transistor, acting as a switch, in series with each device. Indeed, this approach would have been adopted were it not for certain restrictions imposed by the CMOS process.

The process chosen for the fabrication of the ASiTEST1 chip was European Silicon Structures (ES2) ECPD15. This is a 1.5μm, double metal, single polysilicon, 5V digital process: it has been something of a tradition within the Edinburgh Neural Group to construct analogue neural chips using what is essentially a low cost, digital process. The fact that the a-Si:H address circuitry is required to operate with voltage pulses as high as 14V means that any design must consider the substrate diodes in addition to the usual NMOS and PMOS transistors. The substrate diodes associated with the ES2 ECPD15 process, along with their reverse breakdown voltages, are illustrated in figure 3.4.

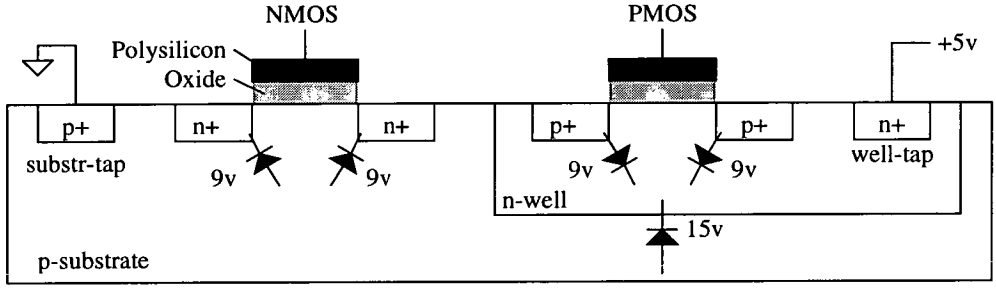


Figure 3.4 - Substrate diodes associated with the ECPD15 process

These diodes impose limits on the voltages which the CMOS transistors can be used to switch. The first set of conditions to consider are those that cause the substrate diodes to be forward biased. This occurs if the drain of a PMOS transistor is raised above 5.7 V or the drain of a NMOS transistor is lowered below -0.7 V. It is this set of restrictions that makes a single transistor select switch impractical. This is best illustrated by considering the situation where an NMOS transistor is used as the select switch for an a-Si:H resistor. The transistor operates as expected during the application of positive pulses, as the drain voltage is between 0 V and +5 V. However, on applying a negative pulse the drain voltage falls below 0 V so the diodes associated with the drains of unaddressed NMOS transistors become forward biased. This causes the programming pulse to appear across all the devices in the array, not just the one selected, as illustrated in figure 3.5.

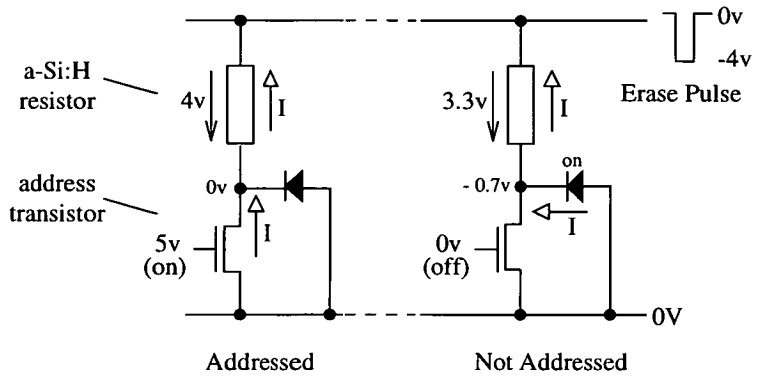


Figure 3.5 - Problem with a single address transistor

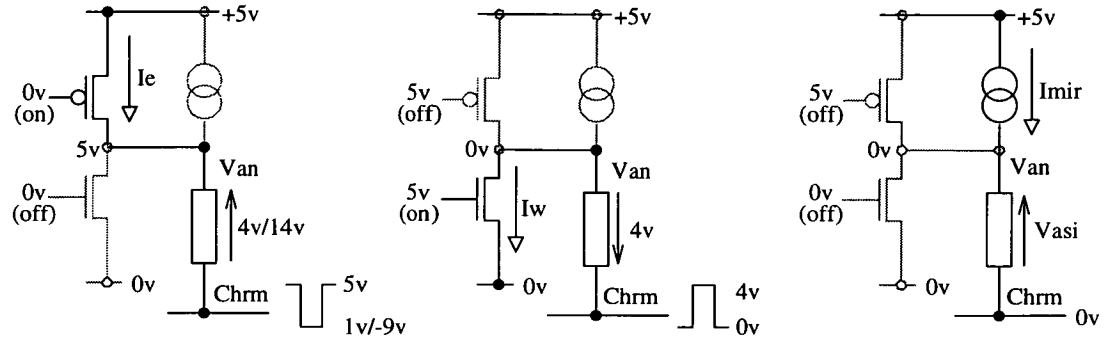
The second set of conditions to consider comprises those that cause the substrate diodes to exceed the reverse bias breakdown voltage. If the drain voltage of an NMOS transistor is raised above +9V then the substrate diodes will break down, again bypassing any address circuitry. This eliminates the practicality of using any form of address circuitry in which transistors are connected directly to the programming rail, as the +14 V forming pulse will exceed the reverse bias breakdown voltage of any diodes connected to it.

The address circuit that was eventually devised is considered in the next section. It is referred to as a FWE cell to denote the fact that it can be used for Forming, Write and Erase operations.

3.2.2.1. The FWE cell

The address circuit designed to allow forming as well as write/erase programming pulses to be applied across an addressed a-Si:H device uses two transistors: a PMOS device for applying positive pulses and an NMOS device for applying negative pulses. Its operation can be summarised as follows:

- Positive Form/Erase pulses - The PMOS address transistor is switched on. This connects the vanadium node to +5 V. The programming pulse, applied to the chromium rail, is negative going from +5 V, as shown in figure 3.6(a).
- Negative Write pulses - The NMOS address transistor is switched on. This connects the vanadium node to 0 V. The programming pulse, applied to the chromium rail, is positive going from 0 V, as shown in figure 3.6(b).



(a) Apply +ve erase/form pulse (b) Apply -ve write pulse (c) Application

Figure 3.6 - The three operating modes of the FWE cell.

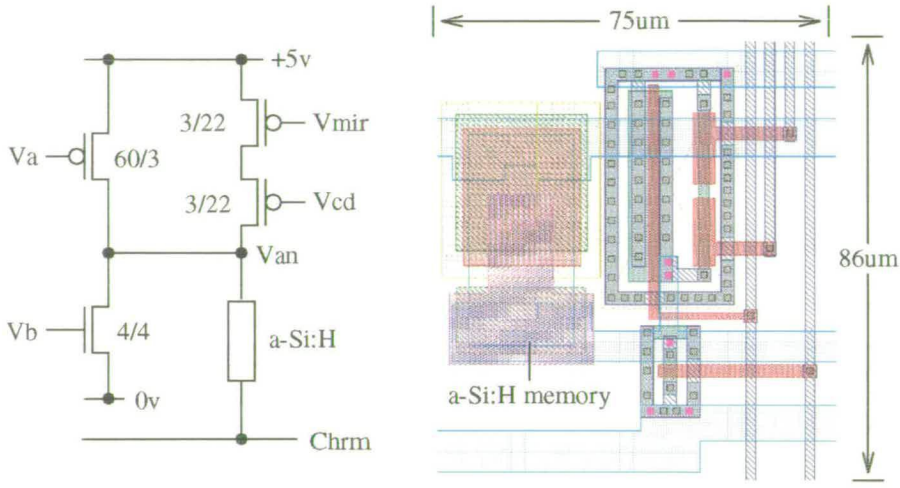
During forming the PMOS transistor holds the vanadium terminal at +5 V. The chromium rail then drops from +5 V to -9 V so that a +14 V pulse appears across the a-Si:H device. By ensuring that the forming pulse is across the a-Si:H device, the address transistor drain remains at 5V, a "safe" digital logic level.

The final feature included in the basic FWE cell is a current mirror, highlighted in figure 3.6(c). This was intended to model a typical application of the memory device, in which the a-Si:H is used to "store" a voltage, determined by the mirror current. Its only reason for inclusion in the FWE cell was to determine whether or not it would affect the programming of the resistor.

The FWE cell was laid out according to the design rules for the ES2 ECPD15 process. The main factor in the layout of the cell was the area required for the a-Si:H memory device. It was decided to use a device 60μm x 40μm with a pore diameter of 10μm. Although this is a very large structure for a device with an active area estimated to be 0.1μm in diameter the size was more a function of the rather primitive photolithography equipment used for a-Si:H processing, rather than any specific device considerations.

The PMOS address transistor was scaled such that its on-state resistance was 1 kΩ, the value of the series resistor used during earlier, resistor array forming. This resulted in a transistor with a W/L ratio of 60/3, a large device. By comparison the NMOS transistor was much smaller, W/L = 4/4, as it was only expected to have to pass voltages much lower than those used during forming.

A schematic diagram of the complete FWE cell, alongside its associated CAD layout, is shown in figure 3.7.



(a) Cell schematic (b) Cell layout

Figure 3.7 - Schematic representation and layout of the FWE test cell

This basic FWE cell was used as a template for the design of three different test circuits, intended to investigate the problem of parasitic programming pulses appearing across non-addressed devices.

3.2.2.2. The three FWE test blocks

In the specification for an a-Si:H address circuit it was stated that to prevent reprogramming, the voltage across a device had to be kept below 0.5 V. During HSPICE simulations of the basic FWE cell it was observed that the drain capacitance associated with the large PMOS address transistor was holding the vanadium node of unaddressed devices at 0 V. This causes a percentage of the programming pulse to appear across non-selected devices, as illustrated in figure 3.8.

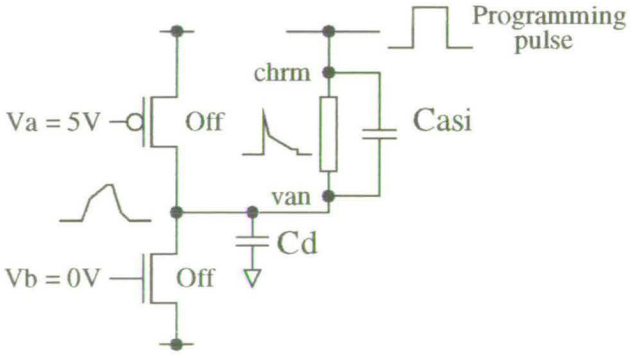


Figure 3.8 - Parasitic programming pulses.

The height of this parasitic programming pulse depends on the ratio of the a-Si:H capacitance, C_{asi}, to the drain capacitance,C_D.

$$V_{asi} = \frac{C_D}{C_{asi} + C_D} V_{pulse} \qquad 3.1$$

For the standard FWE cell layout: $C_D = 0.43pF$, $C_{asi} = 0.1pf$ so $V_{asi} = 0.8V_{pulse}$.

Three variants of this basic cell were therefore designed, each using a different method to minimise the amplitude of the parasitic programming pulse, V_{asi}.

(i) LowC: In this cell the layout of the PMOS address transistor was designed to minimise the size of C_D. This resulted in a C_D value of 0.34pF: $V_{asi} = 0.77V_{pulse}$

(ii) Sanwch: In this design the effective value of C_{asi} was increased by placing a parallel capacitor below the area reserved for the a-Si:H memory. This capacitor is a sandwich of metal2-oxide-metal1-oxide-polysilicon. The additional capacitance $C_{add} = 0.52pF$: $V_{asi} = 0.35V_{pulse}$.

(iii) Driven: In this cell the vanadium node in unaddressed cells was driven to a voltage that prevented parasitic pulses from developing across the a-Si:H device. An example, using a 4V erase pulse, is shown in figure 3.9. The NMOS address transistor in the non-addressed cell has 5 V on its gate. Prior to the pulse its source, connected to Vminus, is also at 5V so the transistor is off. During the pulse Vminus is at 1 V so the transistor turns on ensuring that there is no potential difference across the a-Si:H memory device. The disadvantage of this approach is that two extra control lines, Vplus and Vminus, have to be supplied to the FWE cell.

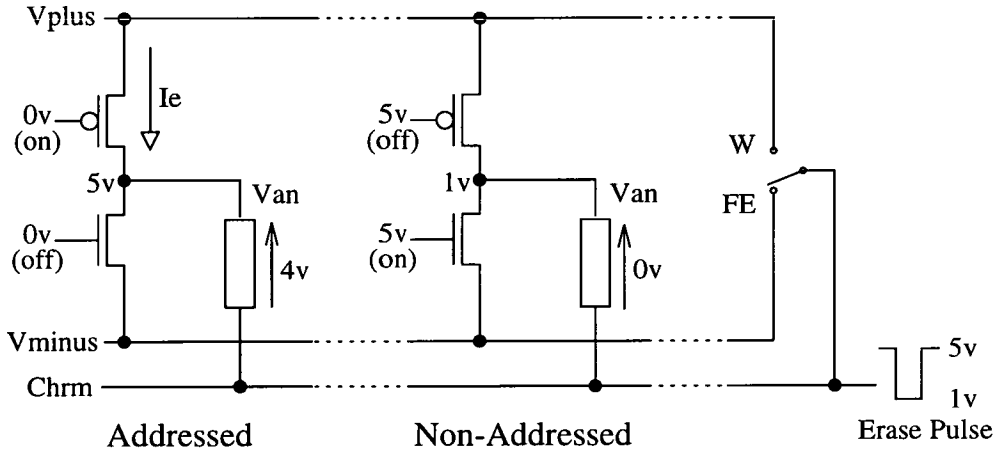


Figure 3.9 - FWE cell where un-addressed nodes are driven to a safe voltage

Each of these three different cell types was the basis of a test block consisting of a 2 x 2 group of the primitive cell. The top row used "Contacts in passivation" memory devices while the bottom row used "Window in passivation" devices.

3.2.3. ASiTEST1 chip overview

As well as the FWE test blocks the ASiTEST1 chip also contained a 4 x 4 array of memory devices without access transistors. The array contained four different memory structures:

(i) 10μ pore memory device: The device used in the FWE cells.

(ii) 5μ pore memory device: A smaller pore than the standard one.

(iii) No pore memory device: An overlap memory, as detailed in Appendix B.

(iv) No a-Si: Short circuit between vanadium and chromium electrodes.

As these two-terminal devices are connected to standard bondpads using metal2 tracks they should highlight any step coverage problems prior to testing on the FWE cells.

A block diagram of the complete ASiTEST1 chip is shown in figure 3.10.

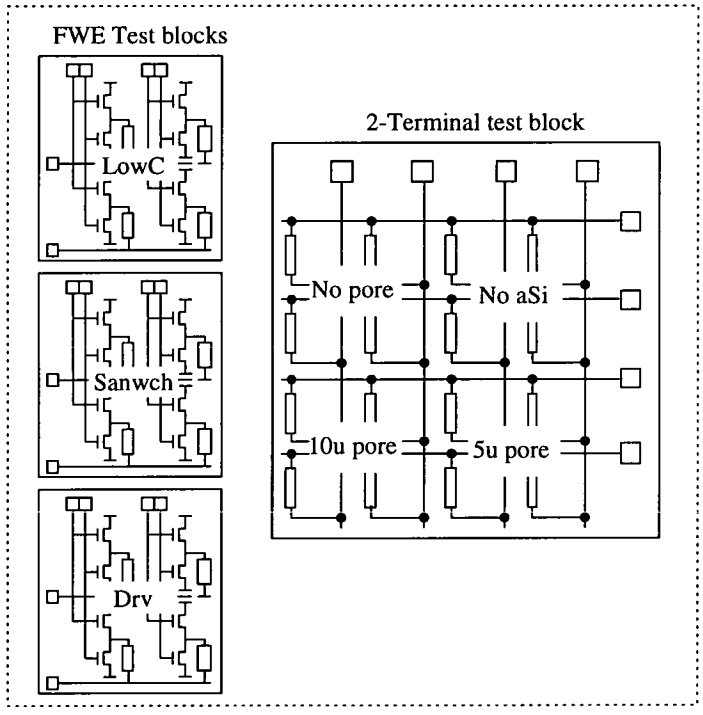


Figure 3.10 - ASiTEST1 test chip block diagram

The FWE test blocks are connected to standard ES2 pads while the array of two terminal devices has its own pads within the chip core.

3.3. ASiTEST1 - Testing

The ASiTEST1 chips were supplied by the manufacturer, ES2, in the form of two half-wafers, instead of the more usual bonded parts. This was to allow the a-Si:H processing to be done more easily on relatively large pieces of wafer. The two half-wafers were cut into a nine different pieces, shown in figure 3.11, in order that various processing experiments could be carried out.

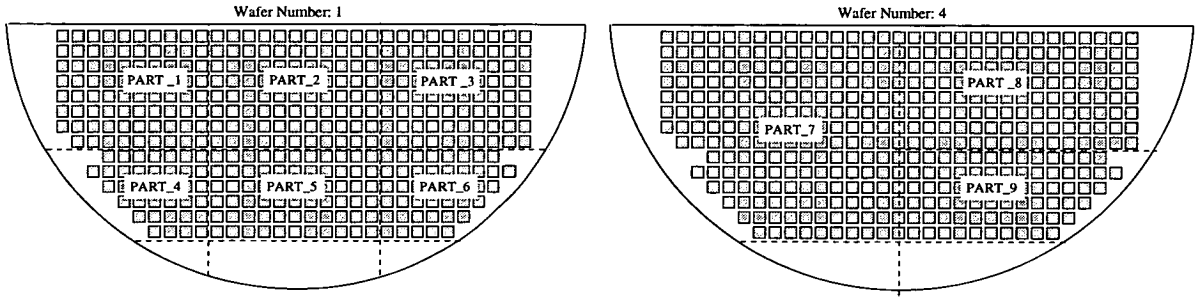


Figure 3.11 - ASiTEST1 wafers

Once a piece of wafer had been processed, a small number of chips were bonded up into either 24 or 40 pin DIL packages:

- 24 pin package - 2 terminal test structures,
- 40 pin package - Three FWE test blocks.

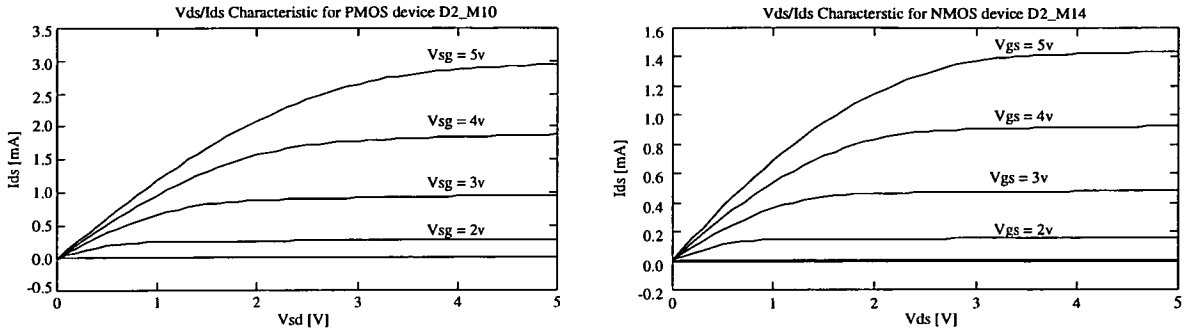
A summary of the various processing and bonding operations carried out on the ASiTEST1 wafers is given in Table 3.1.

Part	Processing	Date	Chips Bonded	Comments
1	All a-Si Layers	4.12.92	24: H2 H3 H4	Working a-Si device.
			40: G4 G8	MOSFETs damaged.
8	No a-Si Layers	14.12.92	40: D1 E1	MOSFETs OK.
5	Sputter Etch and Cr	14.12.92	40: C2	MOSFETs OK.
	Plasma Etch chip C2	27.1.93	40: C2	MOSFETs still OK.
3	All a-Si Layers	27.1.93	24: D1 F1	Working a-Si devices.
			40: C2 D2	MOSFETs OK.
		10.2.93	24: E2 F2 G1 G2	Working a-Si devices.
		17.2.93	24: A3 D3	Working a-Si devices.

Table 3.1 - ASiTEST1 processing and bonding summary

As table 3.1 shows, the first piece of wafer to be processed yielded working a-Si:H memory devices but MOSFETs that had sustained damaged. The MOSFETs were characterised using a Hewlett Packard (HP) parameter analyser. The Vds/Ids curves that were obtained suggested that the transistor's gate oxide had been damaged during the a-Si:H processing.

The next set of wafer segments were therefore tested after each process step that could have potentially caused such damage. By reducing the intensity of the plasma etch operations wafer segments with working a-Si:H devices and undamaged MOSFETs were eventually produced. Figure 3.12 shows the address transistor characteristics from a wafer segment 3 chip that contained a-Si:H devices.



(a) PMOS device characteristic (b) NMOS device characteristic

Figure 3.12 - Vds/Ids characteristics from Wafer_Part_3 devices

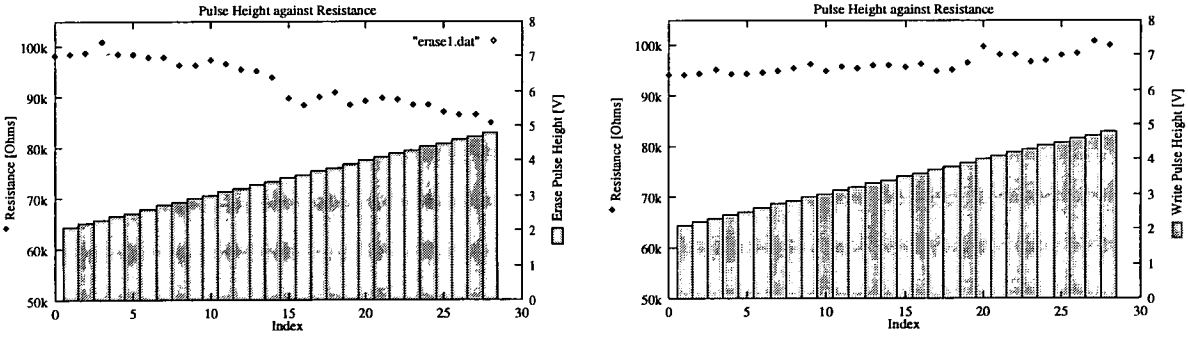
Subsequent testing on the ASiTEST1 chip can be divided into two main sections:

- Testing the FWE test blocks.
- Switching experiments on the 2-terminal test structures.

3.3.1. Testing the FWE cells

In order to supply the control voltages and address signals needed to operate the FWE test blocks a board, detailed in Appendix D, was designed and constructed. The board uses an HP pulse generator to supply programming pulses and a Keithly digital multimeter to record the resistance of the a-Si:H devices.

The following figure shows a series of 120 ns erase/write pulses † (boxes) and the subsequent resistance (points) of the a-Si:H device being programmed. The pulses are displayed in the order that they were applied i.e. there was no reset pulse between each programming pulse.



(a) Device Resistance during a Sweep of Erase Pulse Height (b) Device Resistance during a Sweep of Write Pulse Height

Figure 3.13 - FWE cell switching results

This set of results, in which there is only a relatively small change in resistance with programming pulses up to +5 V, is typical of those obtained from different FWE cells. This inability to produce large changes in the resistance of the a-Si:H device meant that it was

† 120 ns is the width of programming pulse used in all switching experiments, as discussed in section 2.4.3.

impossible to compare the performance of the different designs of programmer cell. It was therefore decided that experiments should be carried out on the 2-terminal test structures to determine the conditions necessary for switching.

3.3.2. Two-terminal switching experiments

The 10μm and 5μm pore devices in the 2-terminal test array on wafer part 3 formed at voltages of between 8 V and 16 V. These devices could then be switched into different states using voltage pulses of less than 5 V. Figure 3.14 shows characteristics from five devices on an ASiTEST1 chip before and after forming.

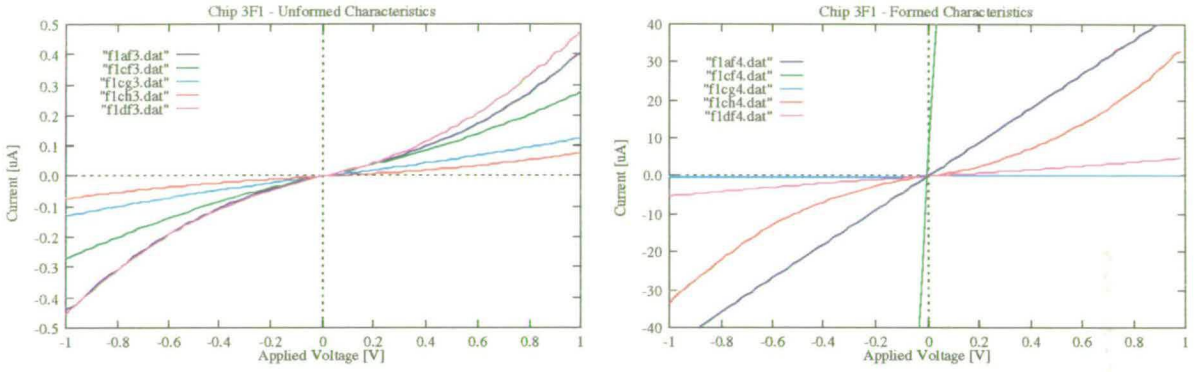


Figure 3.14 - Characteristics of unformed and formed devices

As figure 3.14 demonstrates, the 2-terminal devices, taken from the same wafer segment as the FWE test chips, form into a variety of different resistance states, unlike the earlier MOSFET addressed ones.

These I-V characteristics also demonstrate the non-linearity of the a-Si:H resistor characteristics. It is because of this non-linearity that resistance readings, taken using a digital multi-meter that records the d.c. resistance at 0.3 V, should be used only as an indication of a device's conductivity relative to others, rather than as an exact measure of its behaviour.

The switching experiments attempted on the MOSFET addressed devices were repeated on the two terminal devices. In this case the devices switched over three orders of magnitude from 1 kΩ to 1 MΩ. Figure 3.15 shows a set of programming pulses and the resulting a-Si:H device resistance.

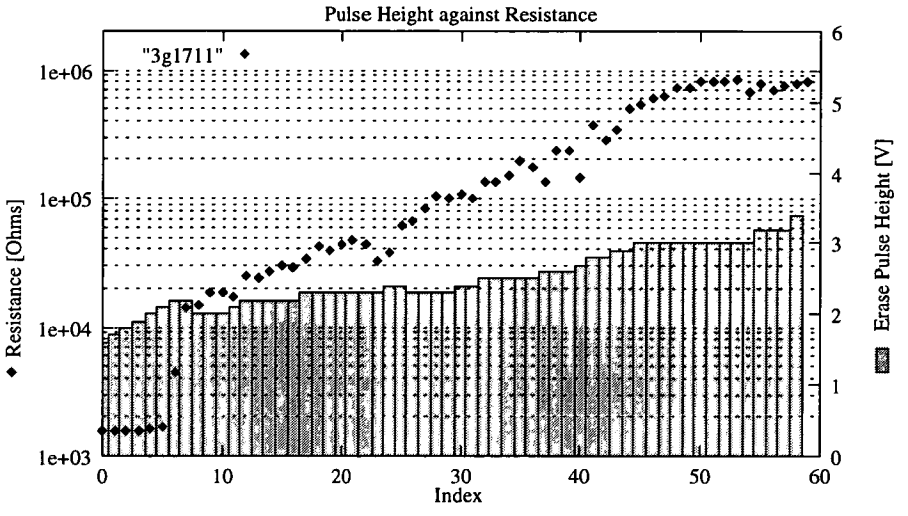


Figure 3.15 - Two terminal switching results - experiment 1

The experimental results displayed in figure 3.15, show that the programming pulse, which was gradually increasing in height, switched the memory device into a large number of different resistance states between 1 kΩ and 1 MΩ. In other experiments it was verified that each new state that a device switched into was quite stable and could be read repeatedly using the multi-meter, without causing reprogramming.

However, many of results obtained during switching experiments were by no means as regular as those shown in figure 3.15. For example, in the following set of results the same programming strategy used in experiment 1 above, was used on an adjacent device in the test array. The pulse heights and corresponding resistances are shown in figure 3.16.

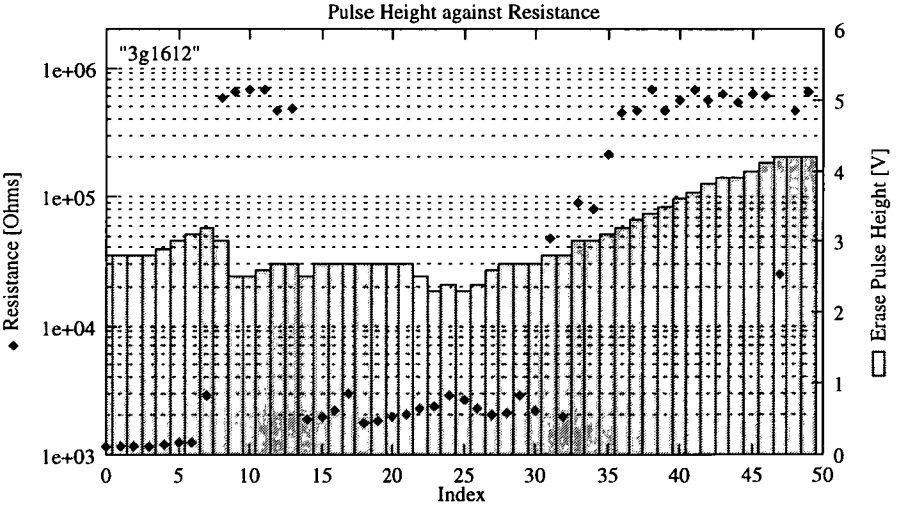


Figure 3.16 - Two terminal switching results - experiment 2

In this case the device resistance jumped directly from 2 kΩ to 700 kΩ instead of through the intermediate states as before. The programming pulse then changed the resistance from 400 kΩ to 2 kΩ i.e. changed resistance in the "wrong" direction. This switching

behaviour is not indicative of a damaged device, as the same device would then switch through a number of different states before reaching 700 kΩ again. These results may reflect an intrinsic instability with a-Si:H devices caused by the uncontrolled forming process.

Another experiment commonly used to characterise the a-Si:H memory device is one in which each programming pulse is followed by a reset pulse. The reset pulse should force the device into the same resistance state each time so that the effect of individual programming pulses of increasing amplitude can be observed. For the results shown in figure 3.17 each erase pulse was followed by a 3 V write pulse, used to reset the device.

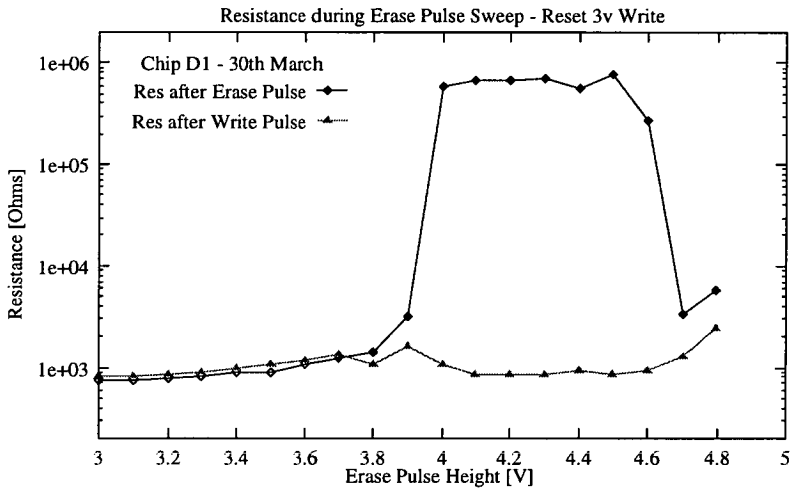


Figure 3.17 - Two terminal switching results - experiment 3

As figure 3.17 shows, each reset pulse results in the memory device changing to a low resistance state, as expected. However, there is no real programming window in which the device can be set to intermediate resistances dependent on the height of the programming pulse, as seen on the original memory devices constructed on glass.

The results of these experiments suggest that a more practical method of programming a-Si:H resistors to a specific value is one in which the magnitude of the applied pulse is increased gradually, rather than one where the amplitude of a single programming pulse is relied upon to give a particular resistance.

3.3.3. A model of switching behaviour

One of the primary objectives of the ASiTEST1 chip was to compare different address circuits intended to reduce the likelihood of devices being re-programmed by parasitic pulses. As the a-Si:H devices associated with these test circuits could not be programmed at all it was decided that a model of device switching, suitable for use in simulations of future address circuits, should be developed using the results of experiments on two terminal devices.

In order to construct such a model, the device behaviour before, and during, switching had to be recorded and analysed. The normal method of generating the I-V characteristic of an unknown device is to use a parameter analyser in voltage sweep mode. However, if voltages of greater than 0.5 V are applied across an a-Si:H device then there is a chance that it may be reprogrammed. This is unfortunate as the region where the device characteristic is most interesting is in the range 2 V to 5 V, where switching occurs. An alternative approach was to use a digital oscilloscope to capture and store the device behaviour.

Using a digitising oscilloscope it was possible to capture and store the waveforms that accompany the 120 ns programming pulses used to change the resistance state of the memory device. Channel A of the oscilloscope was used to display the voltage across the a-Si:H device and Channel B the voltage across a 100 Ω resistor in series with it. In this way the current through the device, as well as the voltage across it, during a programming pulse can be captured. By gradually increasing the amplitude of the programming pulse and recording the scope traces for each pulse it was possible to build up an approximate I-V characteristic for a device in a given resistance state.

The first set of results to be considered comprises the oscilloscope traces up to and during a device transition from 600 kΩ to 784 kΩ. The programming pulses applied to the device and the corresponding device resistances are shown in figure 3.18.

Note: In this figure the write pulses cause an increase in the device's resistance rather than a decrease as is more usual. This is similar to the erase pulses in figure 3.16 also causing resistance changes in the "wrong" direction.

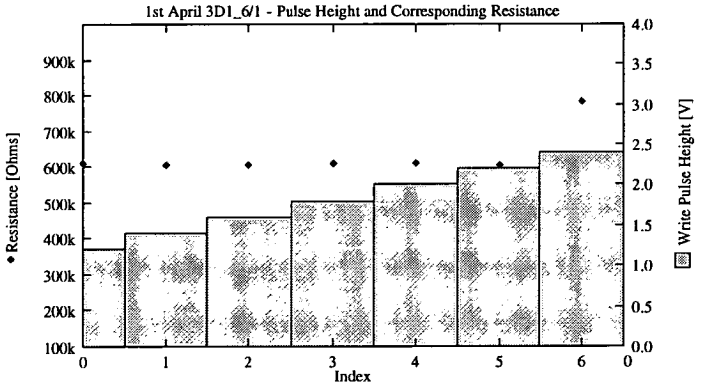


Figure 3.18 - Scope experiment 1: Pulse height and corresponding resistance

The scope traces captured during this series of pulses are shown in figure 3.19. On the final pulse the current through the device increased rapidly and the device ended up in a higher resistance state.

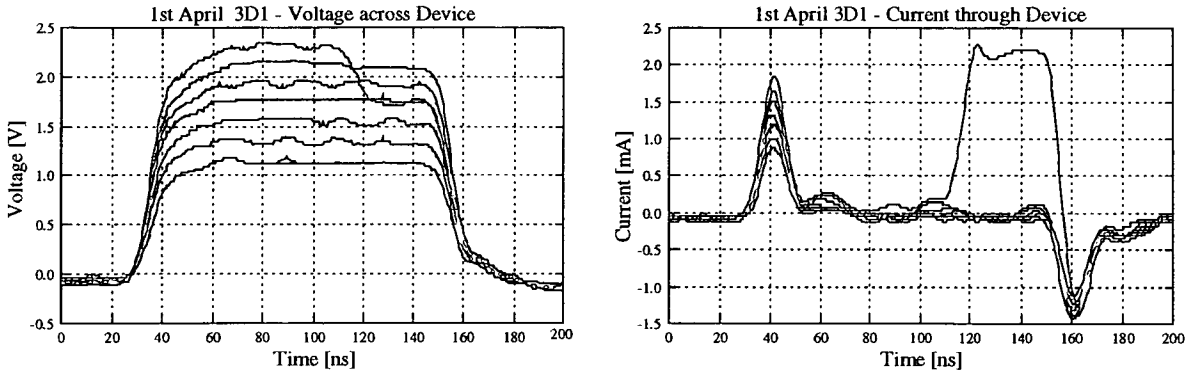


Figure 3.19 - Accumulated set of oscilloscope traces corresponding to the series of programming pulses shown in Figure 3.18

For each pulse prior to the change of resistance the steady state I-V values were recorded from the scope traces. For the final pulse a pair of readings, corresponding to before and after the change, were recorded.

Figure 3.20 shows three sets of I-V points accumulated from oscilloscope traces. In the first set of results, scpa, the measured resistance was 500 kΩ until the final pulse when it dropped to 1.4 kΩ. The second set of results, scpb, show the characteristic of this 1.4 kΩ device until its resistance jumps to 176 kΩ. Interestingly the 1.4 kΩ characteristic coincides with the state change recorded during the 500 kΩ set of results. The final set of results correspond to a device with a resistance of 195 kΩ.

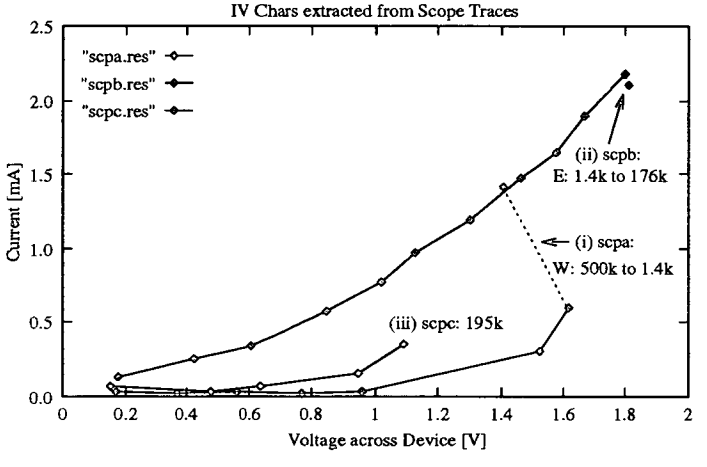


Figure 3.20 : Three I-V characteristics constructed from scope traces

Although the obvious conclusion to be drawn from the first set of switching results is that a change of resistance state is signalled by a high current pulse, this is not always the case. In the following set of results the device resistance again changes during the last voltage pulse, as shown in figure 3.21.

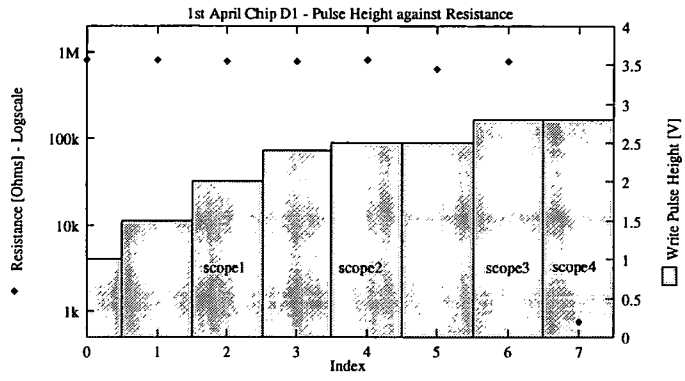


Figure 3.21 - Scope experiment 2: Pulse height and corresponding resistance

However, as the oscilloscope traces of figure 3.22 show, there is a high current pulse in the two scope traces preceding the final one that causes a resistance change.

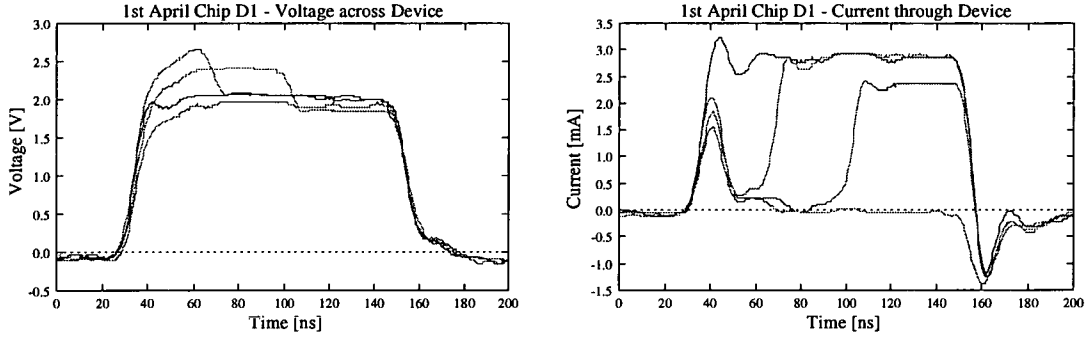
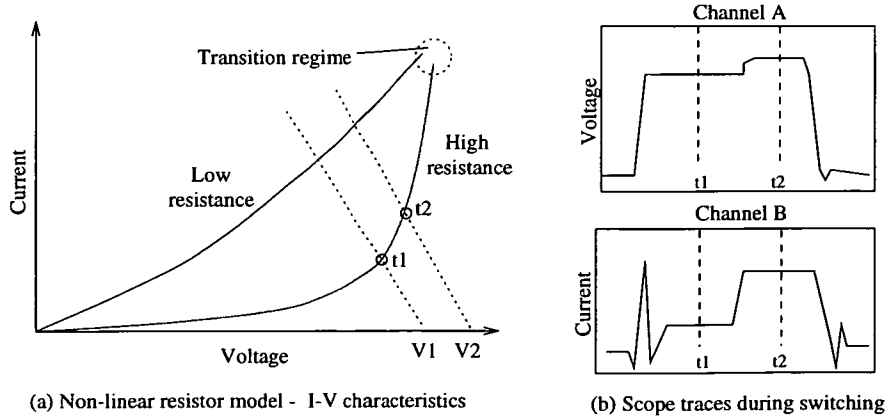


Figure 3.22 - Accumulated set of oscilloscope traces corresponding to the series of programming pulses shown in Figure 3.21

High current pulses do not therefore always signal a change of resistance. This should be reflected in any model of switching behaviour i.e. the high currents should not be caused solely by a resistance change.

The first model that was used in an attempt to simulate this switching behaviour is based on a non-linear resistor. A set of idealised characteristics for a low and high resistance device is shown in figure 3.23(a).



(a) Non-linear resistor model - I-V characteristics

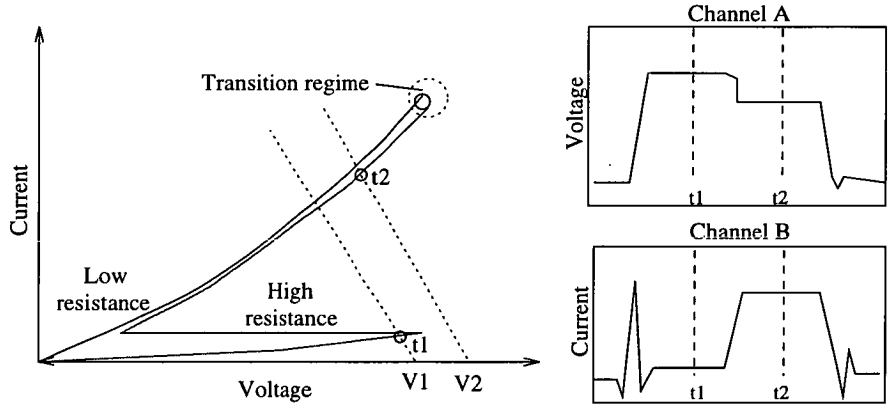
(b) Scope traces during switching

Figure 3.23 - Switching model based on non-linear resistor.

In addition to the device I-V characteristics this figure contains two additional features: firstly, a transition regime where the switching is thought to occur and secondly, two load lines which indicate the stable operating points associated with applied voltages of magnitude V1 and V2.

Figure 3.23(b) depicts the scope traces that the non-linear model produces as the voltage across a high resistance device increases from V1 to V2. If this trace is compared with an actual scope trace then there is an obvious difference. In the captured traces the high current pulse is accompanied by a drop in the voltage across the device, whereas the trace associated with the non-linear resistor model shows a voltage increase.

As the model based on a non-linear resistor was obviously incorrect an alternative, that did produce the correct oscilloscope traces, had to be developed. The new model, shown in figure 3.24(a), assumes that the device resistance collapses above a critical voltage causing the device to enter a high current regime.



(a) Breakdown model - I-V characteristics (b) Scope traces during switching

Figure 3.24 - Switching model based on breakdown.

As figure 3.24(b) shows, the scope trace associated with this model produces a drop in voltage at the same time as the high current pulse. HSPICE models based on this characteristic, and detailed in Appendix E, were used in the design of address circuits on subsequent chips. It was later found that this breakdown characteristic, typical of filamentary conduction, had already been reported in a-Si:H analogue memories by Rose in 1991[77].

These results suggested that it might be more informative if the a-Si:H devices were characterised using the HP parameter analyser in a current sweep mode, rather than in the voltage mode used previously.

For the set of results shown in figure 3.25 a current sweep was used to characterise a device in a number of different resistance states.

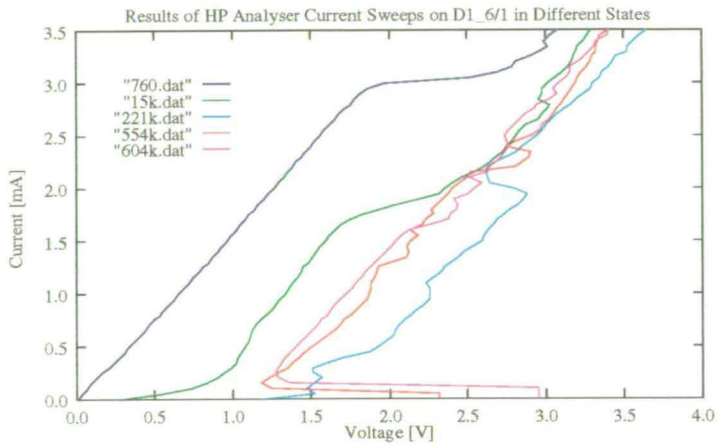


Figure 3.25 - Current sweeps on a device in different states

As figure 3.25 shows, the actual device characteristics are very similar to those used in the construction of the switching behaviour model: the device is stable up to some critical voltage, the resistance then collapses and the device enters a high current regime. It is interesting to note that the device in the lowest resistance state, 760 ohm, behaves as a linear resistor up to 3 mA where its characteristic changes and becomes more like that of the other devices.

It should be noted that the device resistance was invariably changed during the current sweeps that extended in the mA regime. However, this was not the case for sweeps in the μA regime.

3.3.4. Low current operating regime

During the various analyses performed using the current sweep approach it was observed that the device characteristics were quite stable during low current sweeps, 0 to 50μA, even though the voltage across the device rose above 0.5 V, the level previously taken to mark the onset of the programming regime. Figure 3.26 shows a set of characteristics generated using a 0 to 50μA current sweep.

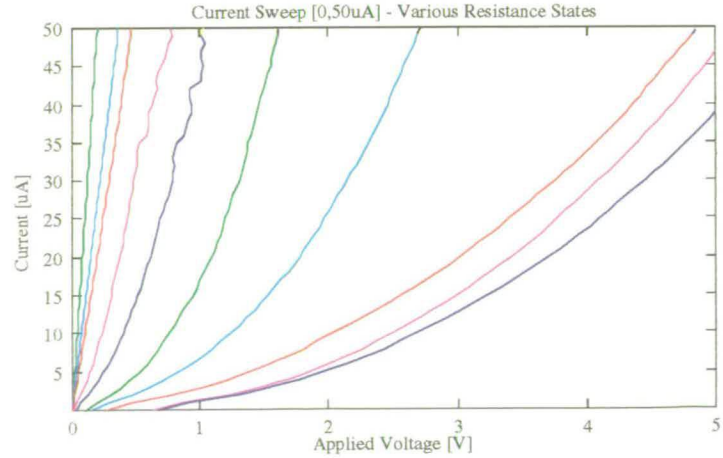


Figure 3.26 - a-Si:H in different resistance states

This result implied that by operating different devices at the same current, say 25 μA, they could be used to store voltages in the range 0 V to +5 V.

3.4. Discussion

The main objective of the ASiTEST1 chip was to show that a-Si:H memory devices could be integrated with CMOS address circuitry. Although the test devices with access transistors could not be switched into different states, the devices in the two terminal test array could. As these devices were connected to the CMOS metal2 layer through holes in the passivation this showed that there were no problems with step coverage and that working devices could be fabricated on the surface of a CMOS wafer.

Experiments on the two-terminal test devices suggested that the reason for the lack of switching in the test cells with access transistors was that the high currents needed for switching were being limited by the performance of the transistors. As the Vds/Ids characteristics of the address transistors showed, the maximum current that could be passed by the PMOS device was 3 mA while the NMOS device could only pass 1.4 mA. The current limiting effect of the transistors would therefore explain the inability to produce a resistance change in the devices associated with access transistors. This effectively solved the problem that the test cells were designed to investigate, namely, how to prevent the programming of non-addressed devices in an array. Memory devices can only change resistance state if the address transistors are turned on, and they can supply sufficient current.

However, the need for high currents during a resistance change means that address cells required larger transistors than those on ASiTEST1 to ensure that devices could change state. The transition at 3 mA displayed during the current sweep on a 760 Ω device, provided a guide to the size of currents that address transistors must be able to pass.

Empirical evidence from the switching experiments carried out on the different two-terminal memory test structures showed that the most robust were the 10μm pore, contact in passivation devices.

The final result obtained from the ASiTEST1 chip was that if the devices were operated below a critical current then they would not be re-programmed. This new operating regime allows memory devices to be used to "store" much higher voltages than the previous limit of 0.5V.

Chapter 4

ASiTEST2 - Synaptic weight storage using a-Si:H

4.1. Introduction

In the previous chapter results from the first test chip, ASiTEST1, were used to demonstrate that working a-Si:H memory devices could be fabricated on the surface of a CMOS chip. The aim of the second test chip, ASiTEST2, was to test various synapse circuits prior to the design of a complete ANN chip, ASiTEST3.

The basic operation to be performed by a synapse cell can be illustrated by again considering the BT XOR demonstrator discussed in chapter 2. The demonstrator uses a chip containing an array of a-Si:H devices to provide the synaptic weight storage for a small ANN, as illustrated in figure 4.1.

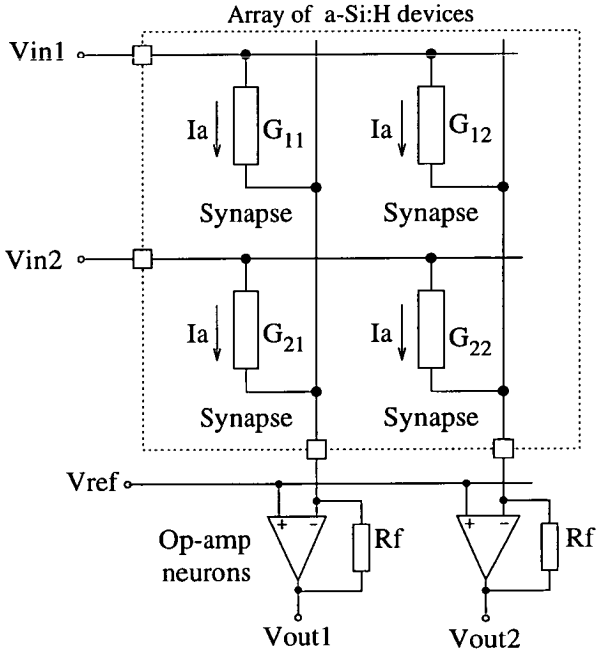


Figure 4.1 - ANN based on an array of a-Si:H memory devices

The circuit operates as follows:

- In the synaptic array the voltage inputs are "multiplied" by the conductance, G_{ij}, of the a-Si:H devices to give a current, Ia.
- These individual currents are then "summed" by the op-amp neuron at the foot of each synaptic column.

- The neuron output voltage, V_{out_j}, is dependent on the summed current and the op-amp feedback resistor, Rf.

The function that the network performs is thus determined by the resistance of the a-Si:H devices. For a network with M inputs and N outputs:

$$V_{out\_j} = -Rf \left(\sum_{i=1}^{M} G_{ij} V_{in\_i} \right) \quad 4.1$$

where V_{out_j} are the neuron outputs, $1 \leq j \leq N$, and V_{in_i} are the input voltages, $1 \leq i \leq M$. This can be compared with the general expression for an ANN:

$$S_j = f\left(\sum_{i=1}^{M} T_{ij} S_i \right) \quad 4.2$$

where S_i and S_j are the input and output states and T_{ij} is the array of synaptic weights. The threshold operator f( ) is usually either a linear function or a non-linear sigmoid.

A network based solely on an array of a-Si:H resistors has two limitations, beyond the need for external neuron and address circuitry:

- It can only implement positive synaptic weights, as there is no "negative" resistor available. Whilst this is sufficient for some networks, such as associative memories[42] where the requirement is for binary weights (0 or 1), most ANNs require both excitatory (positive) and inhibitory (negative) synaptic weights.
- The input signal must be between -0.5 V and 0.5 V to prevent the a-Si:H device from being reprogrammed. This means that additional interface circuitry would be required to use the circuit in conjunction with standard digital logic that uses 0 V and 5 V levels.

In order to address these issues the synapse designs on the ASiTEST2 chip were based on pulsewidth[78] circuits developed for the EPSILON chip[4]. In a pulsewidth neural network the states S_i and S_j are represented by the width of 5 V digital pulses. An illustrative a-Si:H synapse ANN based on this approach is shown in figure 4.2.

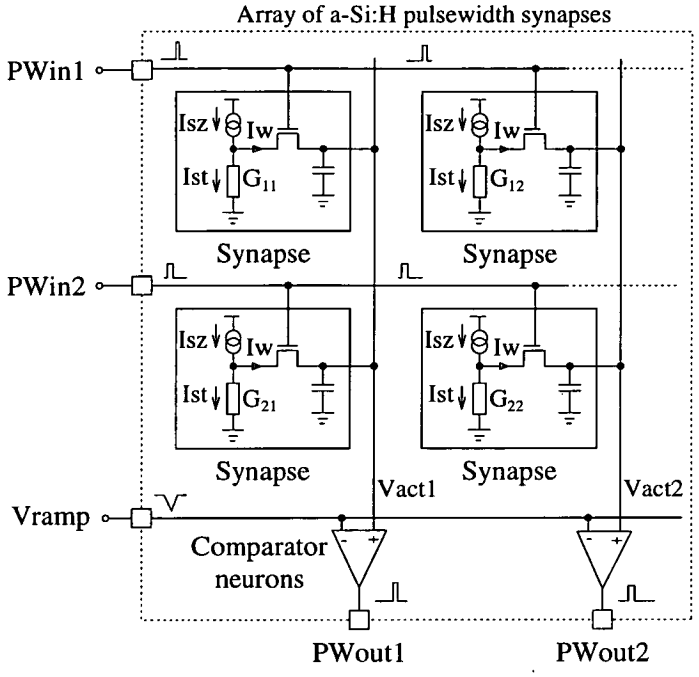


Figure 4.2 - ANN based on a-Si:H pulsewidth synapses

Each synapse cell now contains a current source, Isz, and a pass transistor in addition to the a-Si:H memory device.

- The a-Si:H resistor is used to "store" a current, Ist.
- Ist is subtracted from the "zero" current, Isz, to give a +/- weight current, Iw.
- The weight current, Iw, is gated by the pulsewidth input signal: effectively performing the "input times synaptic weight" multiply operation.
- The current pulses from a synaptic column are summed by the neuron circuitry and the resulting activity is stored on a distributed integration capacitor.

The value stored on the integration capacitor, Vact, is converted to an output pulse using a comparator, the final stage of the neuron circuit. The comparator has two inputs, the activity voltage, Vact, and a ramp signal, Vramp. When the ramp signal falls below the activity level the comparator output goes high. By using a double sided ramp it is possible to generate a pulsewidth output signal, as figure 4.3 shows.

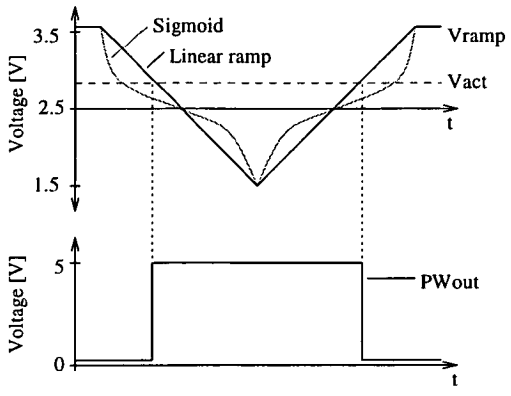


Figure 4.3 - Operation of the pulsewidth comparator

As figure 4.3 also shows, the ramp signal can be altered to implement non-linear thresholding functions, such as sigmoids.

The aim of the ASiTEST2 chip then, was to implement and test various designs of pulsewidth, a-Si:H based, synapses.

4.2. ASiTEST2 - Design

The different synapse designs on the ASiTEST2 chip can subdivided into two main groups distinguished by the type of neuron circuitry: one being the original EPSILON neuron and the other an alternative design suggested by one of the EPSILON designers, but never actually fabricated.

EPSILON's distributed feedback neuron was developed by Donald Baxter to address the issue of process variation in large neural chips[79]. In his circuit, illustrated in figure 4.4, the current pulses produced by the synapses are summed using an op-amp, the feedback buffer of which is distributed amongst the whole synaptic column. The voltage output of this summing op-amp is then compared with a reference signal, Vin_oz, to produce a current that represents the instantaneous activity of all the synapses: if the op-amp output is greater than Vin_oz then the activity current is excitatory, if it is less then the current is inhibitory. This current is then integrated on capacitor Cint to produce a voltage that represents the sum of the synaptic activity.

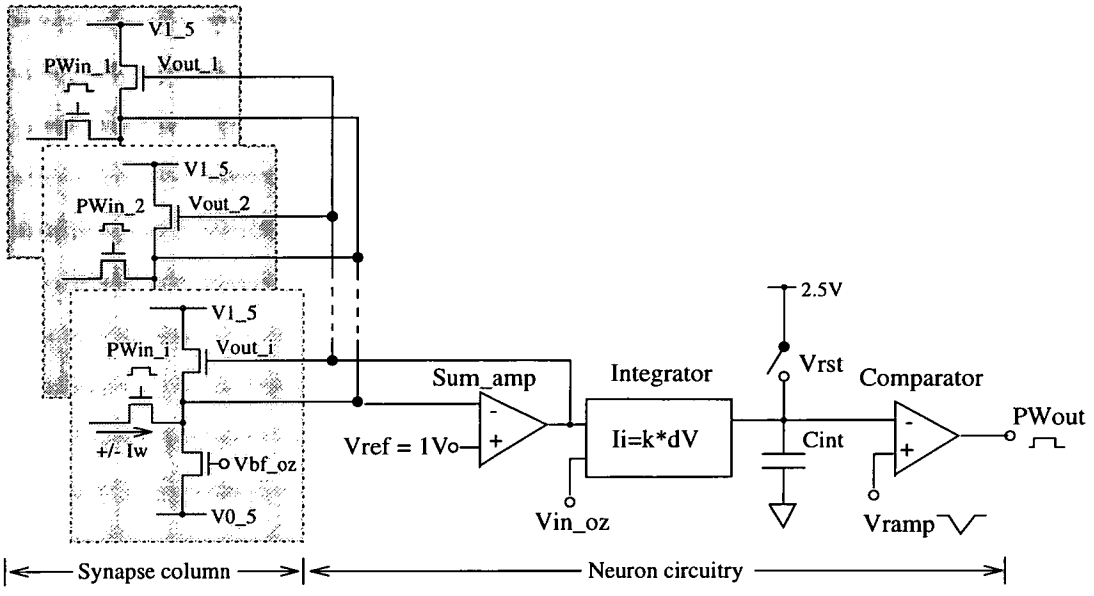


Figure 4.4 - The Epsilon distributed feedback neuron

The other neuron circuit used on ASiTEST2 is based on a design suggested by Steve Churcher in his thesis chapter on pulsewidth circuits[80]. During the course of the project this design was referred to as the Schurch neuron.

In the Schurch neuron the integration capacitor, Cint, is distributed amongst the synapses in a column, as shown in figure 4.5.

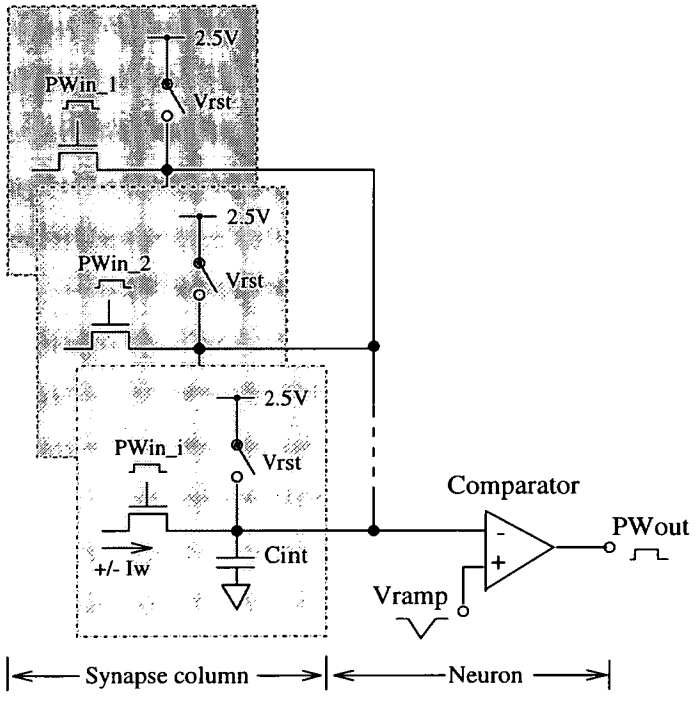


Figure 4.5 - The Schurch distributed capacitance neuron

The current pulses generated by the synapses are now summed directly on the integration capacitor without the need for distributed buffers or integrator circuits. Whilst the individual synapse is probably larger than an EPSILON one, due to the need for an integration

capacitor in each cell, the additional neuron circuitry is reduced to a single comparator.

In the remainder of this section the design of different synapse circuits that produce the weight current, Iw, will be considered. The discussion is divided into the following areas;

- Synapse design issues raised by results from the ASiTEST1 chip
- Design of the EPSILON synapses
- Design of the Schurch synapses
- An overview of the ASiTEST2 chip

4.2.1. Design issues raised by the ASiTEST1 chip

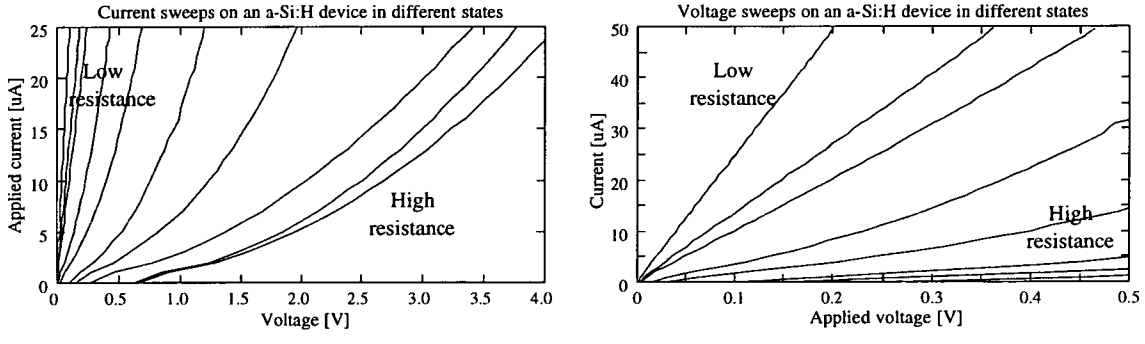
A number of "a-Si:H design rules", based on results from the ASiTEST1 chip, were used in the design of the ASiTEST2 synapses.

(i) Use of the "contact in passivation" approach to integrate a-SiH with CMOS.

(ii) The address transistors should be large enough to supply the 3mA required to switch the device out of the lowest resistance state.

(iii) The high currents needed during state transition mean that additional circuitry to prevent parasitic programming is not required: only devices in cells where the address transistor is on will receive enough current to change resistance state.

(iv) The memory device does not appear to change state so long as the current through it remains below 25μA; this is referred to as the current operating regime, illustrated in figure 4.6(a). The original operating regime was one where the voltage across the device is kept below 0.5 V, as illustrated in figure 4.6(b).



(a) Current operating regime (b) Voltage operating regime

Figure 4.6 - Operating regimes for a-Si:H memory devices

Having defined operating regimes for the memory device its use as the memory element in a synapse cell can now be considered.

4.2.2. EPSILON based synapse designs

In the introduction to this section it was stated that the synapse designs on ASiTEST2 fell into two broad categories. The first set of designs were based on the synapse used on the EPSILON chip, shown in figure 4.7.

The EPSILON synapse [79] is based upon a linear transconductance multiplier. By ensuring that the two transistors Msz and Mst remain in the linear part of their operating regime the currents Isz and Ist will be directly proportional to the gate voltages Vsz and V_{Tij} respectively. It can therefore easily be shown that the weight current, Iw, is equal to:

$$Iw = \beta_T (V_{Tij} - V_{SZ}) V_{DS} \qquad 4.3$$

where β_T is the transconductance of the transistors Mst and Msz.

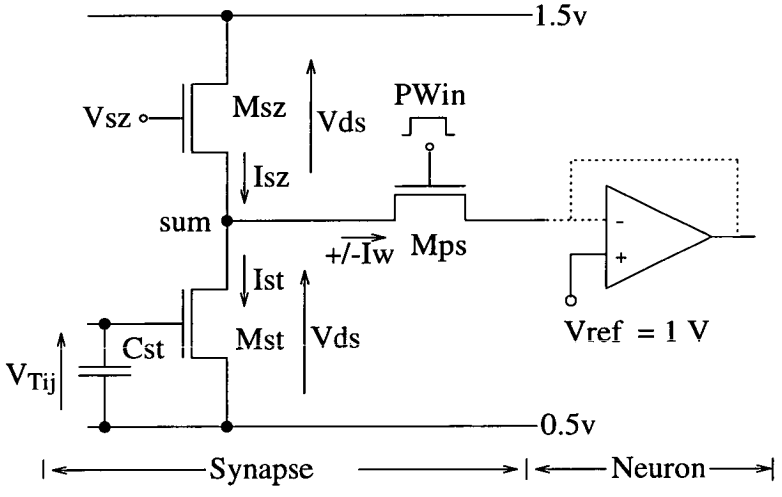


Figure 4.7 - Simplified diagram of the EPSILON synapse

When an input pulse is applied to Mps, the "sum" node is clamped to a virtual 1 V by the action of the summing op-amp. This sets the voltage across both Mst and Msz to 0.5 V, thus satisfying the criteria for linear operation.

The first ASiTEST2 synapse uses an a-Si:H resistor as a direct replacement for Cst, the capacitor used to store the synaptic weight voltage in the EPSILON synapse.

In the "global mirror" synapse a current, Iset, is mirrored to every synapse in the array. The resistance of the a-Si:H memory is then set such that the voltage drop produced by this current gives the same VTij (Vst) as was originally stored using the capacitor. This synapse and the associated Iset cell are shown in figure 4.8.

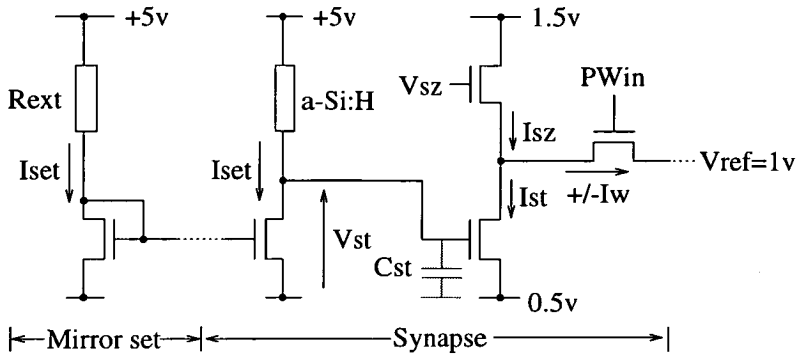


Figure 4.8 - Global mirror synapse schematic

The current Iset is chosen such that the voltage Vst covers the same range as that originally stored using a capacitor. This is illustrated in figure 4.9 which shows a simulated set of a-Si:H characteristics on the same graph as the set current, Iset.

Note: Iset must remain below 25μA to ensure that the a-Si:H remains within the limits of the current operating regime.

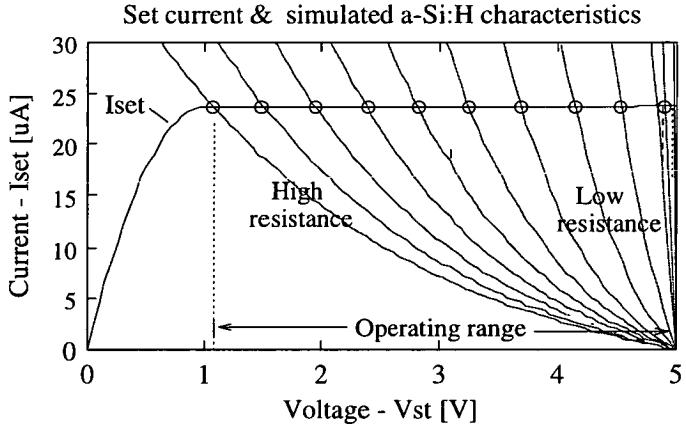
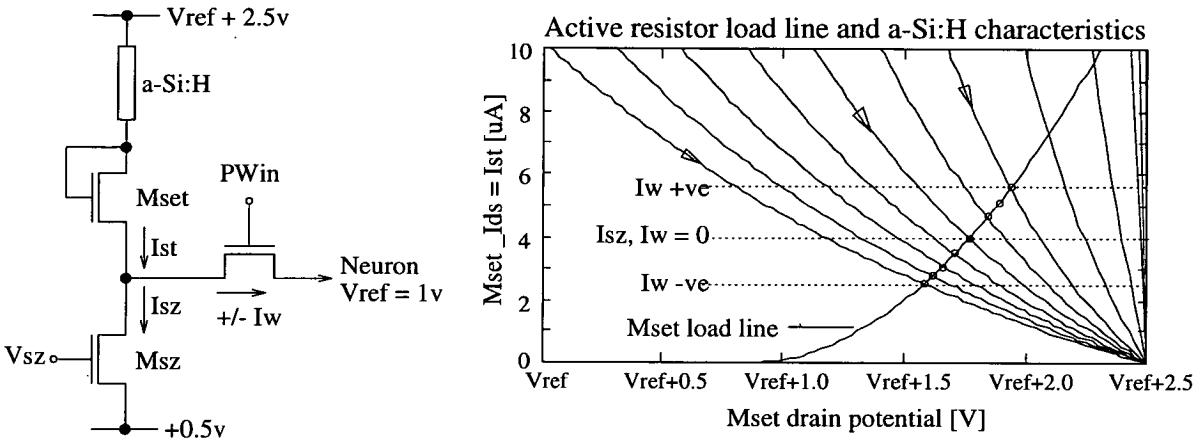


Figure 4.9 - Global mirror synapse operating characteristic

The effects of process variation across a die mean that the Iset current mirrored to each synapse cell will differ slightly from the original. However, during the programming of the a-Si:H device it is the effective weight current, Iw, that is monitored rather than the device's resistance. Therefore, after programming the a-Si:H resistance will be set such that in combination with its copy of the Iset current it produces the desired weight current.

In this design the a-Si:H is used to store a weight voltage which the synapse then converts to a current. The next synapse design uses the a-Si:H to store a weight current directly, thus eliminating the requirement for a globally distributed Iset signal.

In the "active resistor" synapse the a-Si:H is placed in series with a MOSFET being used as an active resistor, as shown in figure 4.10(a).



(a) Active resistor synapse (b) Active resistor synapse operation

Figure 4.10 - Active resistor synapse schematic and operating characteristic

The active resistor limits the current through the a-Si:H device so that a wide range of resistances can be used to store a weight current within the chosen range, 2μA to 6μA. The operation of the active resistor in conjunction with the a-Si:H is illustrated in figure 4.10(b) where the transistor's load line is plotted on the same graph as a set of simulated a-Si:H characteristics.

The variation in the performance of the transistor Mset across different synapse cells means that the resistance needed to exactly match Isz in one cell could result in a positive or negative weight in another. However, the procedure for programming the devices is intended to be an iterative one, in which it is the weight current Iw that is monitored rather than the device resistance. In reality the exact state of the device will be unknown, and only in combination with its Mset transistor will it produce the desired weight current.

The first two synapse designs both use the a-Si:H in its current operating regime. As this regime was based solely on results from the ASiTEST1 chip it was thought prudent to include a third design in which the the a-Si:H device operates within the original voltage regime, that is the voltage across the device is less than 0.5 V.

In the original incarnation of the "constant volt" synapse, the a-Si:H device was simply connected between the sum node and the 0.5 V rail. However, the completed synapse includes a bias transistor in series with the a-Si:H, as shown in figure 4.11.

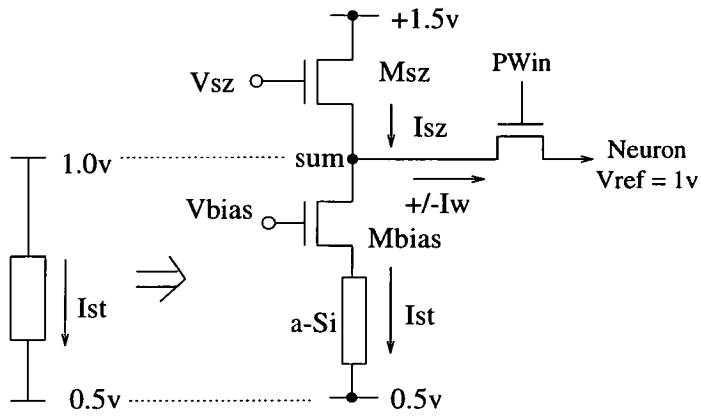


Figure 4.11 - Constant volt synapse schematic

The reason for the bias transistor is probably best illustrated by considering figure 4.12 in which the characteristic of the bias transistor is plotted on the same graph as a set of a-Si:H device characteristics.

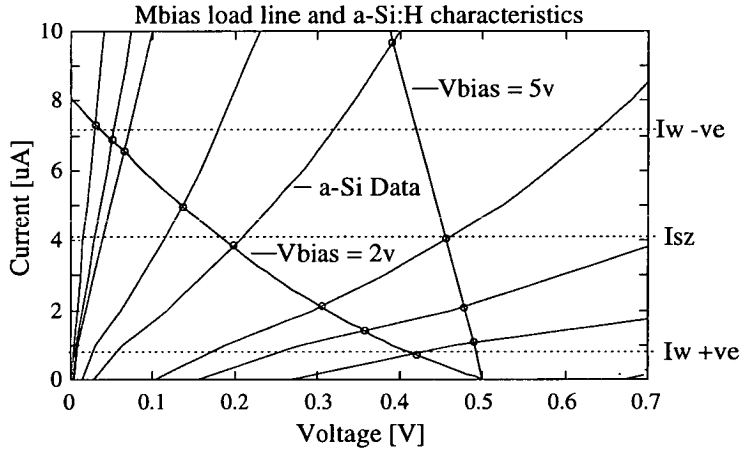


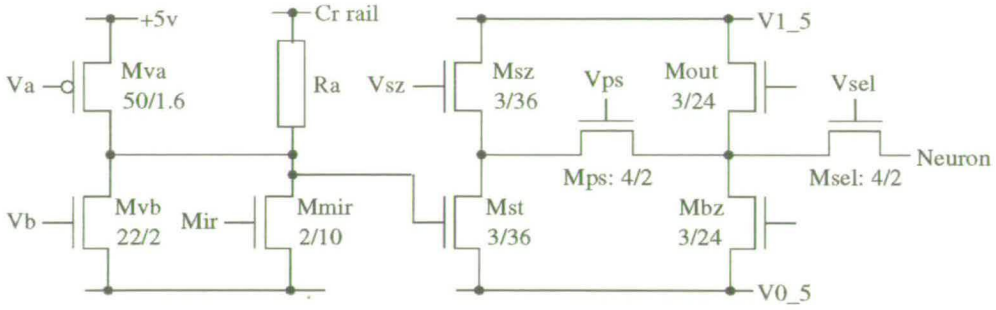
Figure 4.12 - Constant volt synapse operating characteristic

When Vbias is at +5 V the operating regime is equivalent to the original situation where there was no bias transistor: the transistor acts as a short between the memory device and the sum node. In this case the load line only intersects with about half of the available a-Si:H device characteristics. However, when Vbias is lowered to 2 V the increased resistance of Mbias means that the load line intersects all the a-Si:H characteristics in the operating range of 1μA to 7μA.

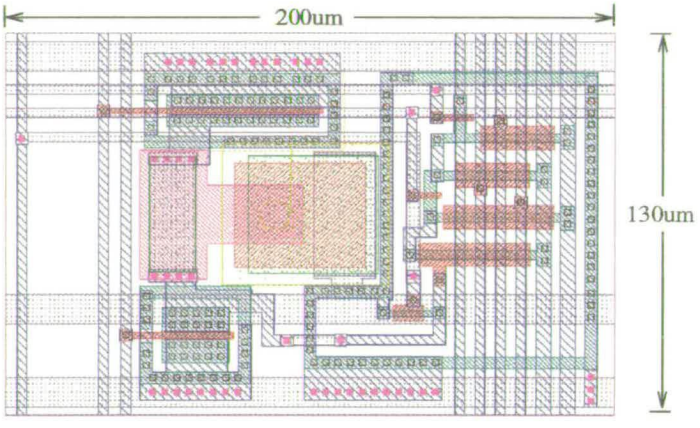
4.2.2.1. Complete EPSILON synapse cell

A complete synapse cell contains a number of extra components in addition to the transistors used to produce the weight current Iw. As all the EPSILON synapses are very similar, only the global mirror synapse will be considered in detail.

A schematic diagram of the complete synapse cell, accompanied by its corresponding layout, is shown in figure 4.13.



(a) Global mirror synapse - schematic



(b) Global mirror synapse - layout

The complete synapse includes the following components:

- The address/programmer transistors Mva and Mvb.
- The local Iset transistor, Mmir.
- The three transistors Msz, Mst and Mps that produce the weight current, Iw.
- The neuron's distributed buffer transistors Mout and Mbz.
- A pass transistor Msel that allows individual synapses to be connected to the neuron circuitry, so allowing the characterisation of single synapse cells.

The next synapse design to be considered is based on the Schurch neuron

4.2.3. Schurch synapse design - Overview

A practical drawback of the EPSILON synapse is the requirement for stable 0.5 V, 1.5 V and 5 V power supplies. In order to overcome this, Churcher designed a pulsewidth synapse that operated from a single +5 V power supply[80]. His original, dynamic storage, synapse is shown in figure 4.14.

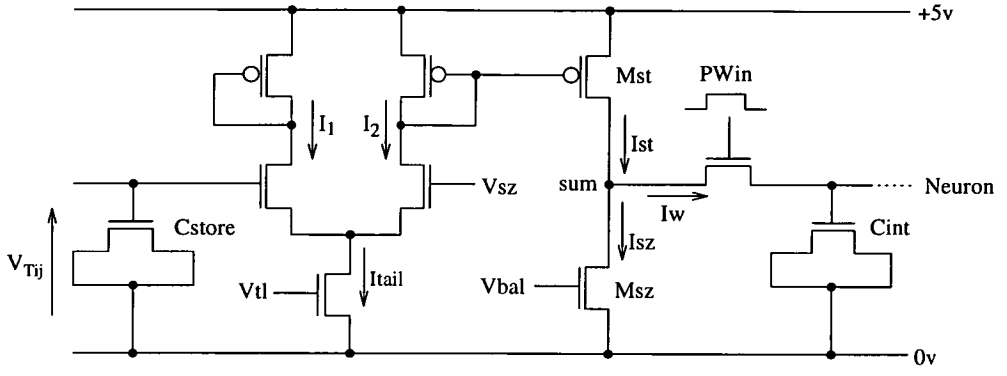


Figure 4.14 - Dynamic storage Schurch synapse

The synaptic weight, V_{Tij}, is held on the capacitor Cstore. It is one input to a low gain differential stage, the other being the zero weight voltage, Vsz. If V_{Tij} is greater than Vsz then I_1 will tend to increase and, as I_1 and I_2 must sum to I_{tail}, I_2 will decrease. This current is mirrored across to Mst where it sums with Isz to give a +/-Iw as on earlier designs.

Note: This dynamic storage version of the Schurch synapse works in an inverting mode. If V_{Tij} is greater than Vsz then charge is actually removed from the integration capacitor, Cint.

Unlike the EPSILON synapse the two transistors Mst and Msz must remain in saturation for the synapse to function correctly. During the input pulse the sum node is at the same voltage as the activity level on the integration capacitor, and not clamped to a constant voltage as in the EPSILON synapse. The activity voltage must therefore remain in the range 1 V to 4 V to ensure that Mst and Msz function correctly - assuming that both transistors have a threshold voltage of 1 V.

While this design was never fabricated, due to funding difficulties, it was decided that an a-Si:H synapse based on this approach should be included on the ASiTEST2 chip.

In the a-Si:H implementation of the Schurch synapse the differential stage, which converts the weight voltage to a current, has been discarded and the resistor is used to store a weight current directly. The basic synapse, along with the zero set cell, is shown in figure 4.15.

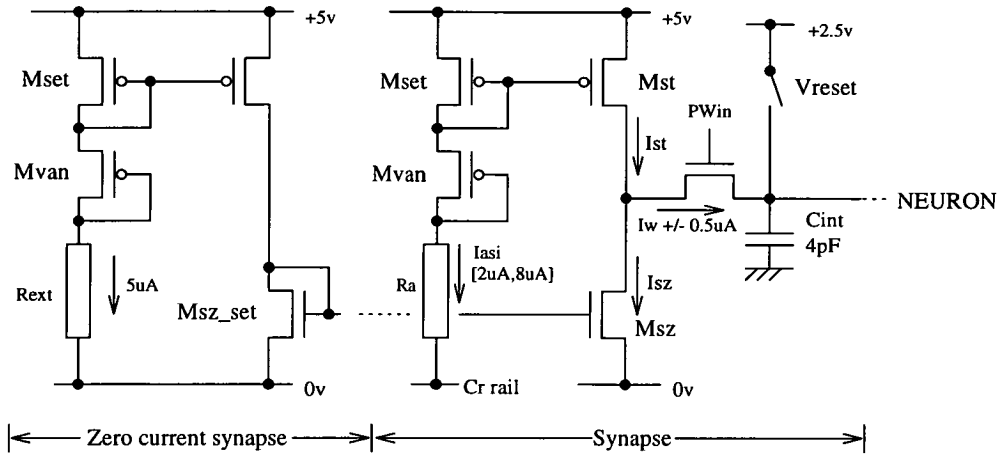


Figure 4.15 - Schurch synapse schematic

The a-Si:H memory is used to store a current, Iasi. This current is mirrored across to Mst where it is subtracted from the zero current, Isz, to give the weight current +/- Iw.

The current Iasi is determined by the resistance of the a-Si:H device and the characteristics of the two transistors Mvan and Mset. In figure 4.16 the characteristic of these two transistors is plotted on the same graph as a set of a-Si:H device characteristics.

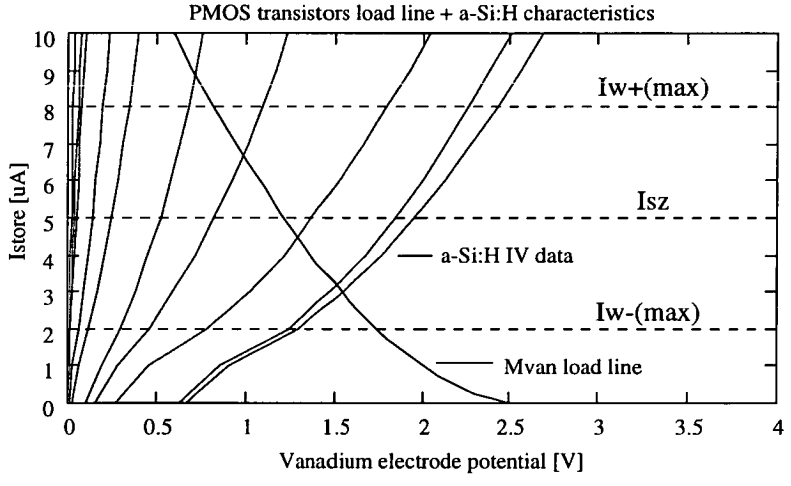


Figure 4.16 - Schurch synapse operating characteristic

The transistor Mvan was chosen such that the transistor load line intersected a large number of the a-Si:H device characteristics, rather like the active resistor design discussed earlier.

Another consideration in the design of this synapse was the magnitude of the weight current Iw. For the circuit to function correctly the voltage on the activity capacitor, Cint, is limited to the range 1 V to 3.5 V; this ensures that the transistors mirroring Isz and Ist remain in saturation. With a reset voltage of 2.25 V the voltage swing that the maximum weight current, in conjunction with the maximum input pulse, should produce is 1.25 V.

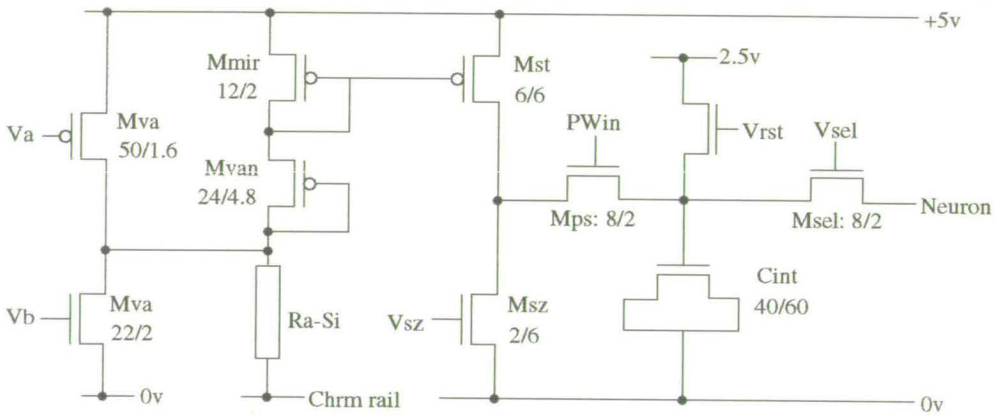
The integration capacitor used has a capacitance of approximately 4pF. The maximum pulsewidth input was chosen to be 10 μs. The maximum Iw current can therefore be calculated as:

$$I_w = \frac{C_{intgn} Vrange}{Tpw} = \frac{4pF * 1.25V}{10\mu s} = 0.5\mu A \qquad 4.4$$

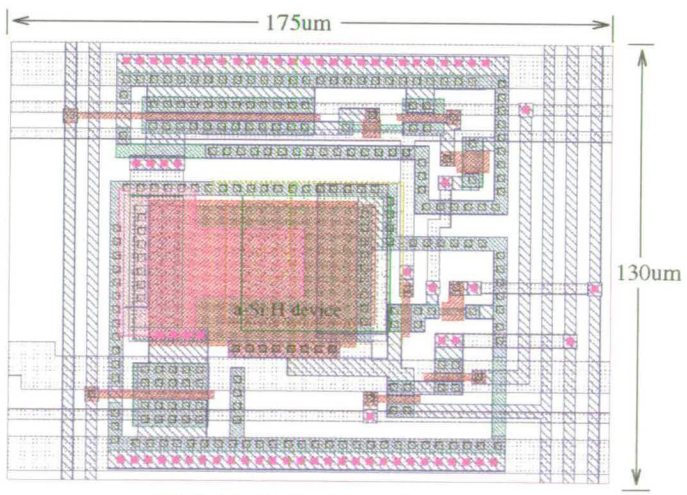
This gives a Iw range of -0.5μA to +0.5μA. However, a useful operating range for the a-Si:H device is 2μA to 8μA. With a zero current of 5μA this gives an operating range of -3μA to +3μA. To scale the current stored by the a-Si:H device by the correct amount the mirror transistor, Mst, is chosen such that its W/L ratio is 1/6th that of the set transistor, Mset.

4.2.3.1. Complete Schurch synapse cell

As with the EPSILON synapses the complete synapse cell contains a number of transistors in addition to the ones needed to produce the weight current.



(a) Schurch synapse - schematic



(b) Schurch synapse - layout

Figure 4.17 - Complete Schurch synapse schematic and layout

The complete synapse contains the following components:

- The programming transistors, Mva and Mvb.
- The five transistors needed to produce the weight current, Iw.
- The integration capacitor with its associated reset transistor.
- A select transistor that allows individual synapses to be characterised.

Two different layout configurations of the Schurch synapse were included on the ASiTEST2 chip. In the first the a-Si:H memory device was laid out over empty, flat substrate, as in the EPSILON based synapses and the two terminal test structures. In the second design the a-Si:H device was placed over the local integration capacitor, Cint, so giving a more compact synapse cell. As the capacitor is a large "flat" structure it should not cause any disturbance in the passivation surface in the region where the a-Si:H device is fabricated. The complete schematic and layout of this second synapse is shown in figure 4.17.

The various synapse designs are included on the ASiTEST2 chip in five test blocks, each containing a column of four synapses.

4.2.4. ASiTEST2 chip - Overview

The complete ASiTEST2 chip consists of two main blocks:

(i) Synapse test block: This contains five columns, each of four synapses, with the associated neuron circuitry at the foot of each column.

(ii) Discrete a-Si:H memory devices: Two terminal test structures of various types.

A block diagram of the ASiTEST2 chip is shown in figure 4.18.

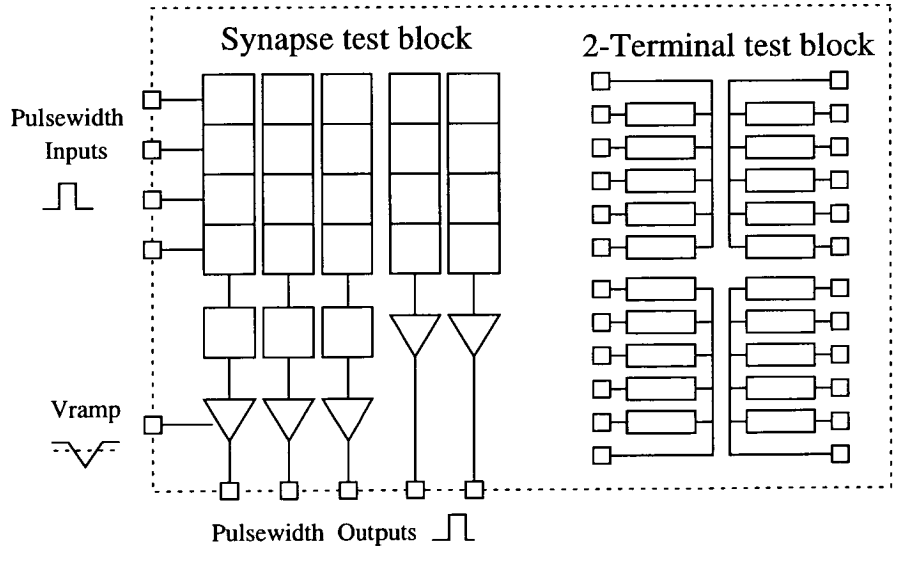


Figure 4.18 - Chip2 Block Diagram

As with ASiTEST1 the main CMOS test cells are connected to standard ES2 pads while the 2-terminal test block has its own pads within the chip core.

4.3. ASiTEST2 - Test system

While the ASiTEST2 chips were being fabricated, two test boards were designed and constructed: one for the EPSILON synapses and the other for the Schurch ones. As well as providing all the synapse and neuron control signals the boards could also be used to apply programming voltages and read back resistance states.

The device to be programmed is selected using the address transistors in combination with the synapse cell's chromium rail. As the address transistor gate voltages, Va and Vb, are common to a whole column of synapses a particular device is selected by closing one of four relays, each of which is connected to a different chromium rail, as shown in figure 4.19.

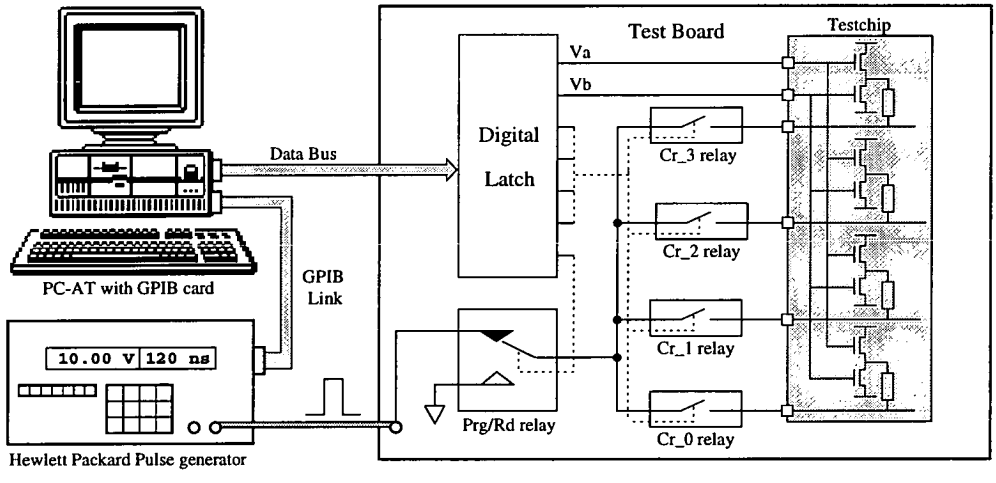


Figure 4.19 - ASiTEST2 test board: Device addressing

The programming pulse is supplied by a HP pulse generator under GPIB control.

To measure the effective resistance of a programmed device the synapse circuit itself is used, as figure 4.20 illustrates. A single buffer stage is selected, using Vsel, so that the neuron output only depends on the contribution of the chosen synapse. A pulsewidth input of fixed duration is then applied to the synapse cell. The charge that is dumped on the integration capacitor, Cint, will depend on the value of the weight current which itself is a function of the a-Si:H resistance.

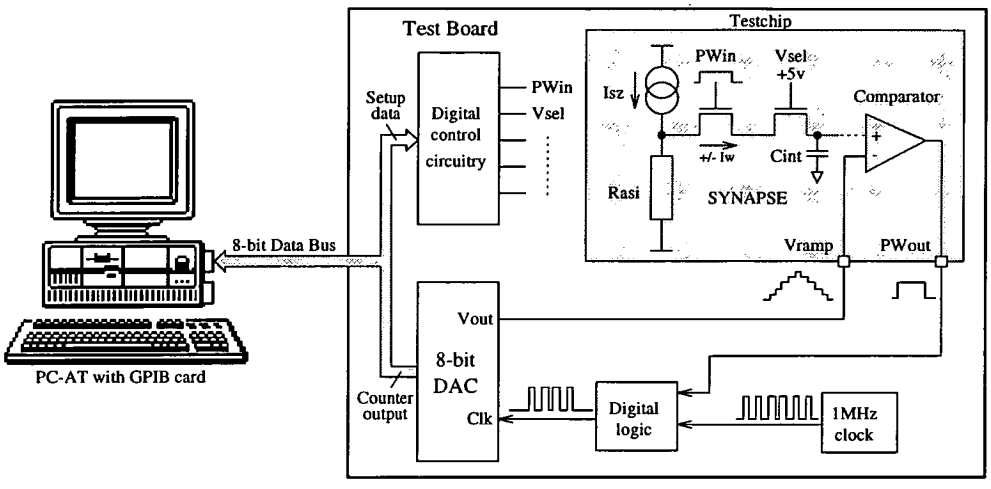


Figure 4.20 - ASiTEST2 test board: "Resistance" measurement

The voltage on the integration capacitor is read using the ramp signal, Vramp, produced by a DAC with an internal counter. When the ramp signal exceeds the voltage on the integration capacitor the comparator output goes high. This transition is used to inhibit the clock pulses supplied to the DAC. The value of the DAC counter can then be read back into the PC and the activity voltage calculated.

By using this method it is the effective weight current, Iw, that is monitored, rather than the resistance of the a-Si:H device.

4.4. ASiTEST2 - Results

The ASiTEST2 chips were supplied by ES2 (July 1993) in the form of three whole wafers. These were cut in half or quarter pieces for a-Si:H processing. After a wafer segment had been processed a small number of die were bonded into DIL packages. As the largest package that was readily available was a 40 pin DIL a total of six different bonding configurations were required to test the ASiTEST2 chip fully. These different configurations are detailed in Appendix C.

Table 4.1 contains a summary of the different ASiTEST2 processing runs and experiments that were carried out.

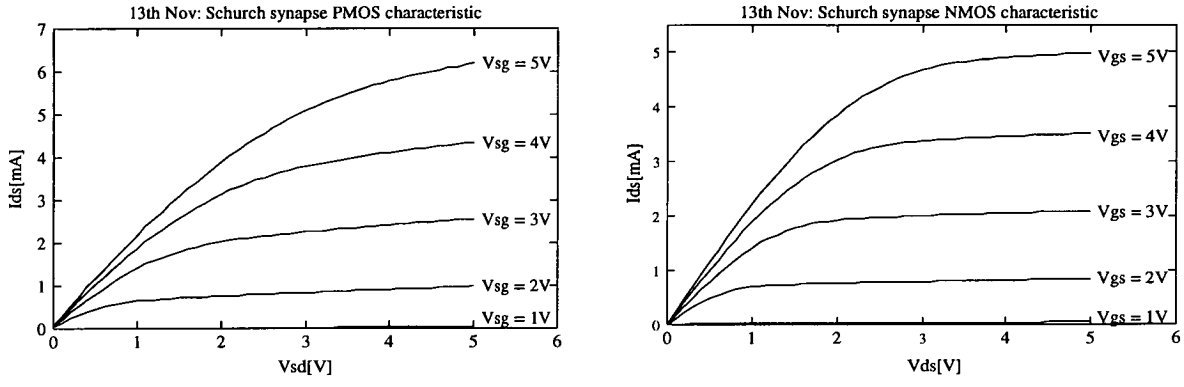
Processing	Date	Comment
No a-Si layers	28.7.93	Wafers arrive from ES2
No a-Si layers	28.9.93	Characterise Schurch synapse with variable resistor
No a-Si layers	5.10.93	Characterise EPSILON synapse with variable resistor
No a-Si layers	13.11.93	Characterise MOSFETS
550A a-Si	17.11.93	Characterise MOSFETS Initial a-Si switching experiments
1000A a-Si	9.3.94	Vanadium transfer layer, very high forming voltages
750A a-Si	9.3.94	Lower forming voltage, switching expts, stability test.

Table 4.1 - ASiTEST2 processing and testing summary

As table 4.1 shows, a number of different experiments were carried out on the ASiTEST2 chips, both with and without a-Si:H. While the a-Si:H layers were being deposited it was possible to characterise both the address transistors and the synapse cell itself using chips that had not yet been processed.

4.4.1. Pre a-Si:H deposition experiments

In order to compare the new address transistors with those on ASiTEST1 they were characterised using the HP parameter analyser as before. The Vds/Ids characteristic obtained is shown in figure 4.21.



(a) PMOS address transitor - Vds/Ids characteristic

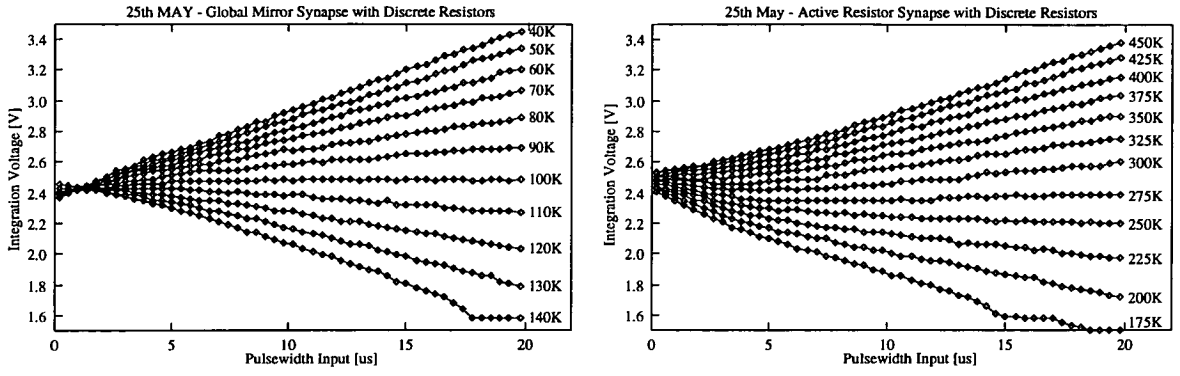
(b) NMOS address transitor - Vds/Ids characteristic

Figure 4.21 - ASiTEST2 Chip address transistor characteristics

As the results show, the address transistors are now capable of supplying currents well in excess of the 3 mA specification.

The other set of experiments carried out prior to a-Si:H deposition tested the operation of the different synapse cells. By using a variable resistor in place of the a-Si:H device it was possible to generate the multiply characteristic of a synapse cell. For different resistor settings the pulsewidth input was swept from zero to maximum width and at each step the corresponding voltage on the integration capacitor recorded.

The first set of results, figure 4.22, show the multiply characteristic of a global mirror and an active resistor synapse.

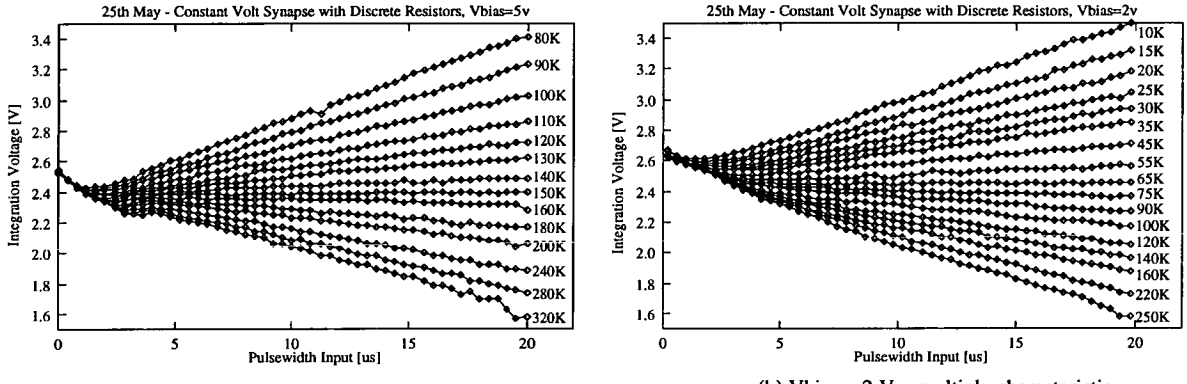


(a) Global mirror synapse - multiply characteristic

(b) Active resistor synapse - multiply characteristic

Figure 4.22 - Epsilon based synapses characterised using a variable resistor

The next set of results, figure 4.23, show the multiply characteristic of a constant volt synapse with Vbias at 5V and then at 2V. The second sweep confirms that the resistance range is increased when Vbias is lowered to 2V.



(a) Vbias = 5 V - multiply characteristic

(b) Vbias = 2 V - multiply characteristic

Figure 4.23 - Constant volt synapse characterised using a variable resistor

As each of the three EPSILON based synapse designs has a reasonable multiply characteristic any decision on the one most suitable as the basis of a complete neural chip would have to be based on some other design criterion. One obvious consideration for a large neural chip is the power dissipated in each synapse cell. On this basis the constant volt synapse, in which the voltage across each a-Si:H resistor is only 0.5 V, would appear to be the most suitable for large arrays.

The last synapse to be characterised using a variable resistor was a Schurch type. As the results produced by the Sch1 and Sch1b were indistinguishable only those of a Sch1b synapse are shown in figure 4.24.

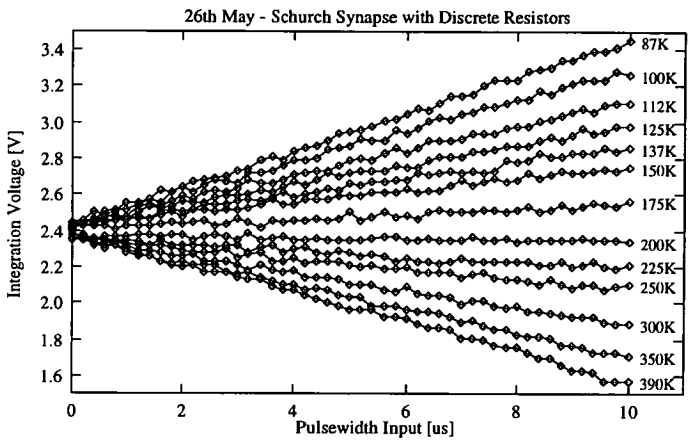


Figure 4.24 - Schurch synapse characterised using a variable resistor

Having characterised each of the different synapse designs using a variable resistor the next stage was to test the synapses on chips with a-Si:H memory devices.

4.4.2. a-Si:H Programmability and Stability

During experiments on synapses incorporating a-Si:H memory devices there is no direct method of measuring the device resistance, which therefore has to be inferred from the results of a synapse calculation. In the results that follow the device "resistance" is represented by the voltage on the integration capacitor following a 10μs input pulsewidth.

In order to demonstrate the switching behaviour found on ASiTEST2 two sets of typical results, taken during sweeps of erase and write pulse heights, now follow. In figure 4.25 the amplitude of a 120 ns write pulse is increased from 2.5 V to 14.5 V.

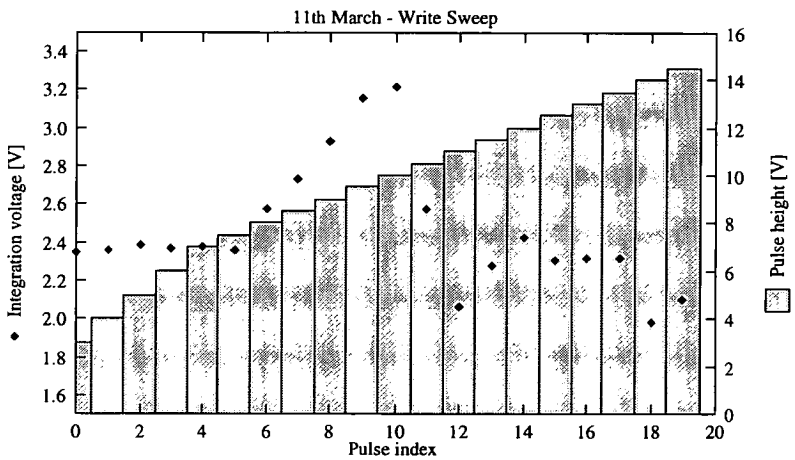


Figure 4.25 - ASiTEST2 sweep of write pulse height

As these results show there was no device switching until the pulse height was about 7 V, much higher than that needed to produce switching on the ASiTEST1 2-terminal devices. The activity level then increased with pulse height indicating that the resistance was decreasing, as it should for write pulses. However, above 10 V the activity level then

drops even though the pulse polarity has not changed.

In the second set of results the erase pulse height is increased from 3 V to 11 V. The pulse heights and corresponding resistances are shown in figure 4.26.

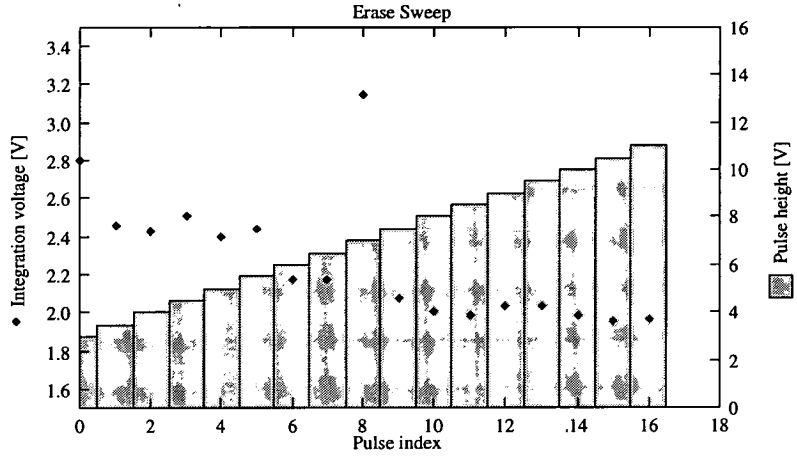
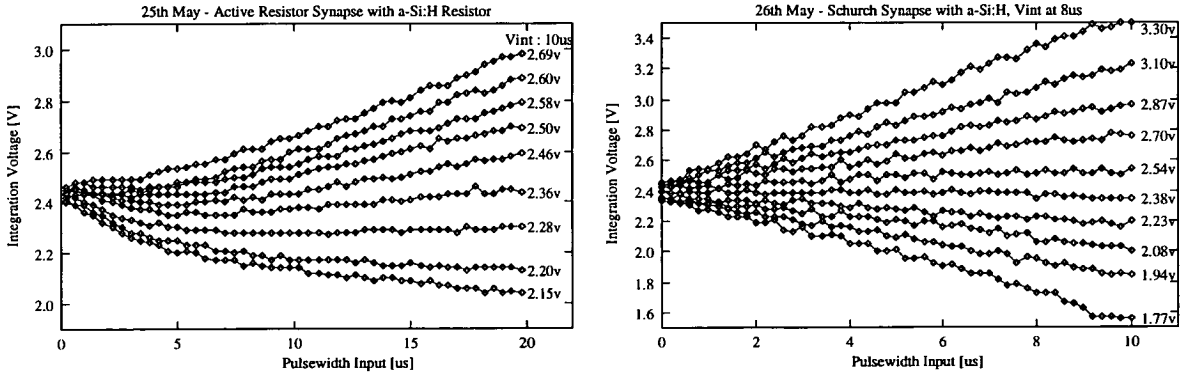


Figure 4.26 - ASiTEST2 sweep of erase pulse height

In this set of results it can be seen that the activity voltage decreased as the pulse height increased. However, once again there was no change of resistance until the pulses were about 7 V in height.

Due to the difficulty of programming the a-Si:H devices on this wafer segment only two sets of a-Si:H synapses characteristics are included here. The multiply characteristics shown in figure 4.27 were taken from an active resistor and an Schurch synapse with a-Si:H memory devices. Due to the occasionally erratic switching behaviour of the a-Si:H device the procedure for generating these characteristics was to carry out a pulsewidth input sweep each time the device was programmed into a previously unrecorded resistance state.



(a) Active resistor synapse - multiply characteristic

(b) Schurch synapse - multiply characteristic

Figure 4.27 - a-Si:H synapse multiply characteristics

As figure 4.27 shows the performance of the two synapses is the same with a-Si:H as with the variable resistor. This is not surprising as the only difference here is the mechanism used to generate the weight current, Iw. However, it does demonstrate that the a-Si:H

devices are stable, at least for short periods, within the current operating regime, where the voltage across the device is much higher than the original 0.5 V limit.

In another experiment the four synapses in a synaptic column were programmed into different resistance states. On each morning for the week that followed, the same chip was powered up for 30 minutes and the state of the devices recorded. Figure 4.28 shows the accumulated results from this experiment.

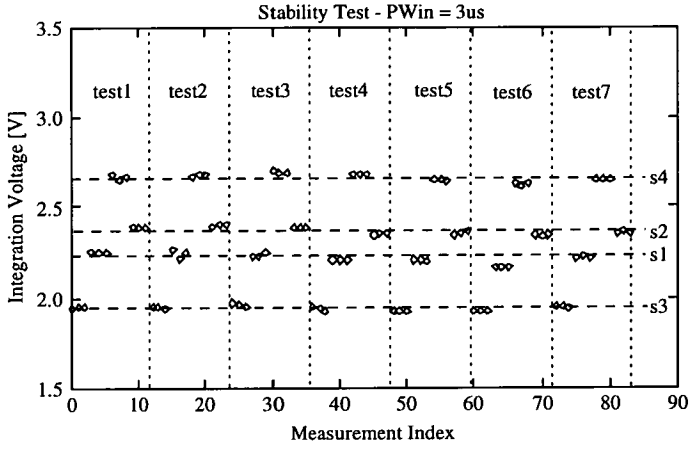
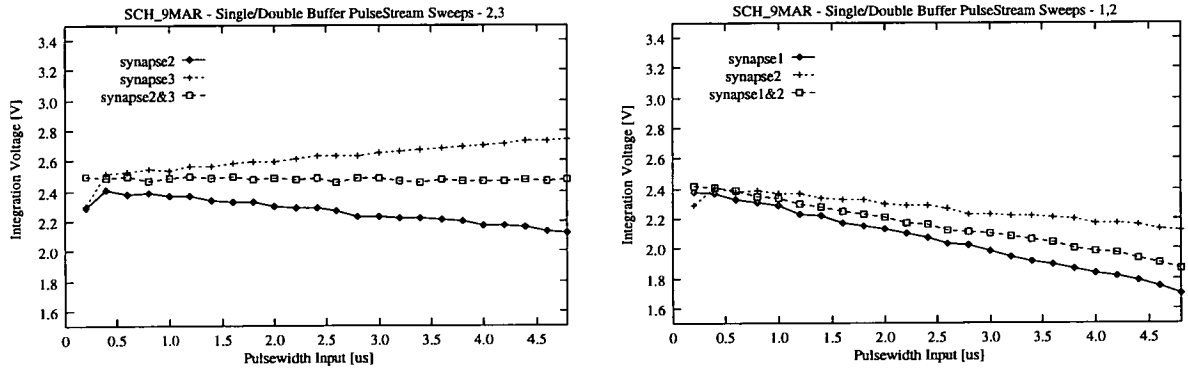


Figure 4.28 - Stability test

As the figure shows the devices remained in the same resistance state over the week long period.

As all the experiments thus far have used individual synapse cells the final one tests the operation when two are switched in simultaneously. Two synapses were characterised individually and then jointly with both Vsel switches closed. As figure 4.29 shows the result is an average of the two individual sweeps as one would expect.



(a) Synapses 2 and 3

(b) Synapses 1 and 2

Figure 4.29 - Single and double synapse sweeps

4.5. Discussion

Due to the need to design and submit the final test chip, ASiTEST3, there was only a minimal amount of a-Si:H based testing done on the ASiTEST2 chip. Once the basic operation of the synapses had been confirmed a decision had to be made on the most suitable design to include on ASiTEST3. The relative merits of the different synapses are discussed in chapter 5.

During testing it was noticeable how difficult the EPSILON based chips were to setup and test in comparison with the Schurch designs. For each new EPSILON based chip the voltages Vin_oz and Vsz had to be manually readjusted in order to obtain a sensible synapse characteristic. Figure 4.30 shows the multiply characteristics taken using two global mirror synapses on different chips without the usual board recalibration.

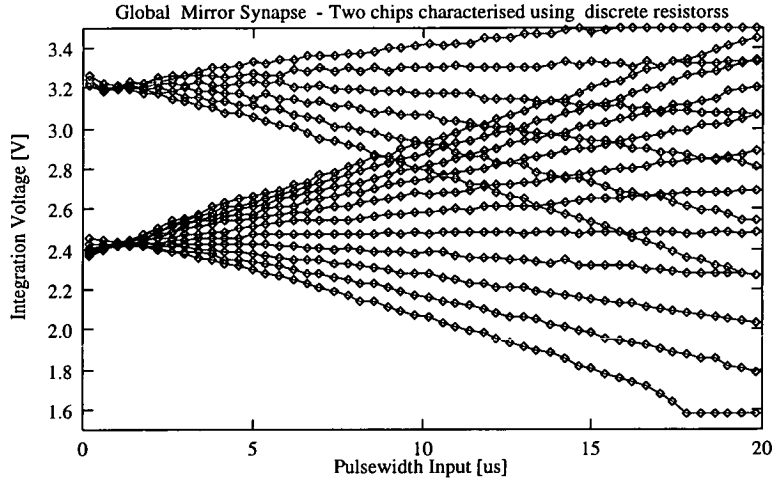


Figure 4.30 - PWin sweeps on two EPSILON chips

The final comment on the ASiTEST2 testing is that although the switching behaviour of the a-Si:H devices was again poor it was sufficient to confirm two things. Firstly, devices could be programmed into resistance states spanning the whole weight space. Secondly, once a device had been programmed it remained in the same resistance state both over the period needed to generate a multiply characteristic and then over a week of periodic testing.

Chapter 5

ASiTEST3- An 8x8 ANN with a-Si:H Synapses

5.1. Introduction

The aim of the final test chip, ASiTEST3, was to implement a small ANN based on one of the synapse designs tested on ASiTEST2. In addition, the chip was intended as a vehicle for exploration of some of the system level issues associated with pulsestream neural chips.

This concern with system level considerations meant that the design of ASiTEST3 had to be such that the chip did not require external programmer relays or a plethora of analogue control signals. This criterion was used to decide which of the ASiTEST2 synapse circuits should be chosen. Table 5.1 compares an EPSILON synapse and an Schurch one from a support circuitry standpoint.

The EPSILON synapse		
Power supplies	Voltage references	Current references
V5_0 = 5.0V	Vin_oz = 3.74V	Ineu = 4μA
V1_5 = 1.5V	Vsz = 3.75V	Iamp = 45μA
V0_5 = 0.5V	Vbf_oz = 3.1V	Iint1 = 2μA
	Vrstv = 2.5V	Iint2 = 2μA
	Vref = 1.0V	

The Schurch synapse		
Power supplies	Voltage references	Current references
V5_0 = 5.0V	Vrstv = 2.5V	Ineu = 4μA,Isz = 5μA

Table 5.1 - System level comparison of the ASiTEST2 synapses

As table 5.1 shows, the EPSILON synapse requires a large number of reference signals as well as three power supplies, all of which must be extremely stable for the transconductance multiplier to function correctly. By comparison the Schurch synapse can be operated from a single 5 V power supply and requires only two reference currents.

The sensitivity of the EPSILON synapse to the value of these references was highlighted by the need for board re-calibration each time a new chip was tested. It was therefore decided to base the ASiTEST3 chip on the Schurch synapse, the EPSILON approach

being more suited to large neural arrays where the cost of system level complexity may be balanced against the increased performance offered by process invariant circuits.

As part of the system level design it was decided that the test board for the ASiTEST3 chip should be designed such that two ANN chips could be cascaded in series, forming a two layer network. On the test boards designed for the dynamic storage EPSILON chip there is only one neural chip. A two layer network is implemented by downloading the weights corresponding to the first layer, performing a forward pass, storing the result, and then downloading the weights corresponding to the second layer. This approach is obviously impractical for a chip that uses non-volatile storage.

This chapter is divided into three main sections:

- ASiTEST3 chip design
- ASiTEST3 test system
- ASiTEST3 results and discussion

5.2. ASiTEST3 - Design

In the introduction to this chapter it was stated that the aim of the ASiTEST3 chip was to construct a neural network chip based on the Schurch synapse. It was decided to base the chip on an 8 x 8 array of synapses, an extremely modest network by neural standards. This array size was chosen for two main reasons:

i) Having eight inputs and eight outputs makes interfacing to digital test circuitry relatively straightforward. For example, a single octal buffer chip can be used to isolate all the chip's pulsewidth inputs from a common data bus.

ii) A chip based on an 8 x 8 array is the largest that can still be bonded into a 40 pin carrier, the largest package readily available.

It was also hoped that by keeping the design this small, the testing would be relatively simple, and the results obtained over a number of chips could still be used to study the feasibility of larger networks based on this a-Si:H memory technology.

A block diagram of the complete ASiTEST3 chip is shown in figure 5.1.

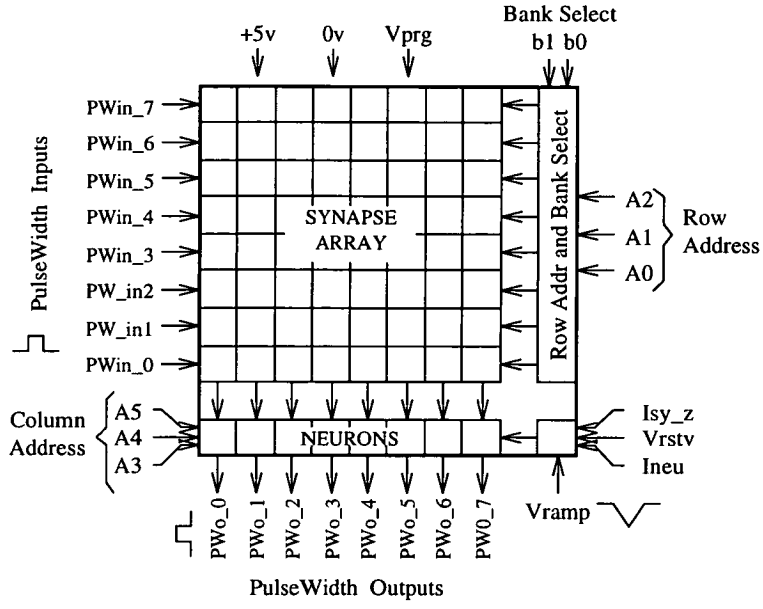


Figure 5.1 - ASiTEST3 chip block diagram

The chip has two operating modes: programming and recall.

In programming mode the synaptic weight to be altered is selected using a 6-bit digital address. An external pulse generator is then used to supply a programming pulse of suitable amplitude and duration.

In recall mode the chip functions as a single layer, feedforward neural network. The eight input pulsewidth states are "multiplied" by the synaptic array to give an activity voltage at the foot of each column. These activity voltages are converted to output pulsewidth signals using a ramp signal, as discussed in chapter 4.

As figure 5.1 shows, the chip can be divided into three main blocks:

- The neuron and column address decoder
- The row address decoder with bank select
- The synaptic array

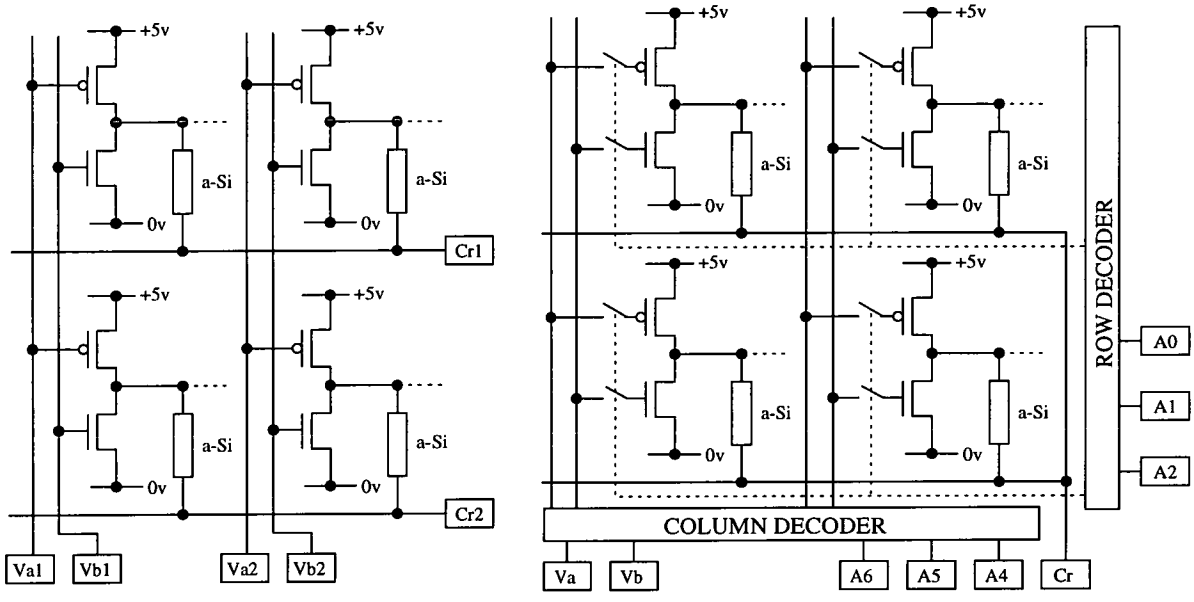
Each of these three main blocks will now be briefly discussed in turn.

5.2.1. Addressing circuitry

One issue to be addressed in the ASiTEST3 design was system level simplicity. The goal of a chip that was easy to set-up and test, combined with a restriction of using a maximum of 40 pads, meant that the addressing scheme used on the ASiTEST2 chip had to be improved.

On the ASiTEST2 chip the device to be programmed was selected using a combination of the address transistor gate voltages, Va and Vb, and the chromium rail associated with the chosen synapse cell; the chromium rail connects together the bottom electrodes of the a-

Si:H devices in a given row. This scheme of using Va/Vb to address columns and using the chromium rail to address rows is illustrated in figure 5.2(a). If this scheme had been used on the ASiTEST3 chip then a total of 24 pads would have been required to address the 8 x 8 array: sixteen for the Va/Vb voltages and eight for the chromium rails. It would also have required external relays to select the different chromium rails and digital logic to generate the Va/Vb voltages.



(a) ASiTEST2 addressing (b) ASiTEST3 addressing

Figure 5.2 - Comparison of ASiTEST2 and ASiTEST3 addressing

The arrangement used on the ASiTEST3 chip requires nine pads: one each for Va, Vb and the chromium rail and the remainder to provide a six bit digital address of the device to be programmed. This arrangement with digital row/column addressing is illustrated in figure 5.2(b). As well as requiring fewer pads, it also eliminates the need for external programming relays, considerably simplifying the design of any interface or test board.

5.2.2. Column decoder and neuron

The column decoder is used, in conjunction with the row decoder, to select the a-Si:H device to be programmed. As there are eight synaptic columns, a 3 to 8 line decoder is used to select the one to which the address transistor gate voltages, Va and Vb, are to be passed, as shown in figure 5.3.

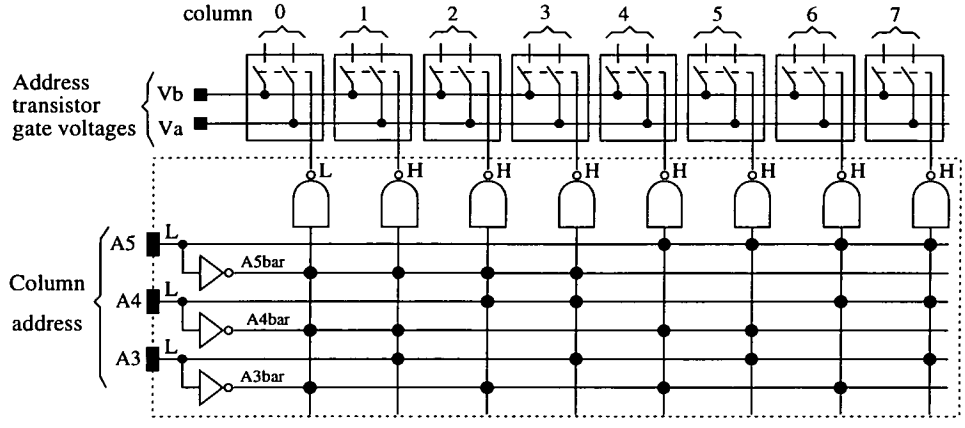


Figure 5.3 - Column decoder schematic

The need for high programming currents during a-Si:H device switching means that in this addresser scheme both the synapse cell and the column decoder must be able to pass full logic levels of +5V and 0V to both Va and Vb. If single pass transistors were used, then the associated threshold voltage drop, giving, say, 1V instead of 0V, would result in the transistors passing currents below their potential maximum. It was therefore decided that transmission gates should be used to pass the address transistor gate voltages, Va and Vb, to the desired synapse cell in the array. Another advantage of using transmission gates is that analogue gate voltages can be supplied to the address transistors, offering alternatives to the existing voltage pulse based programming methods. The transmission gate cell is shown in figure 5.4.

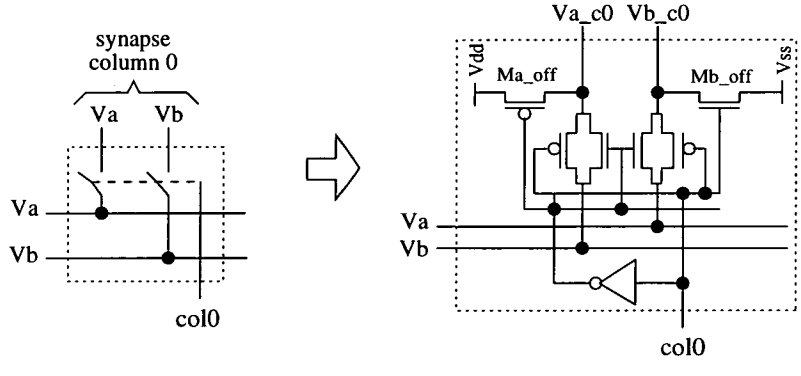


Figure 5.4 - Transmission gate schematic

When the column select line is low, the voltages Va and Vb are switched into the chosen column. However, if the column select line is high, the two additional transistors, Ma_off and Mb_off, ensure that the two address transistors are kept in their off state.

The 3 to 8 line decoder is built from eight instances of a basic decoder cell. The functionality is determined by an additional set of instances which connect the NAND gate inputs either to address or to address_bar lines. The original intention was that the basic decoder cell should be dimensioned such that it was pitchmatched to the width of the synaptic

column. However, by compacting the layout, it was possible also to include the output comparator neuron in the same cell. The layout of this decoder+neuron cell is shown in figure 5.5.

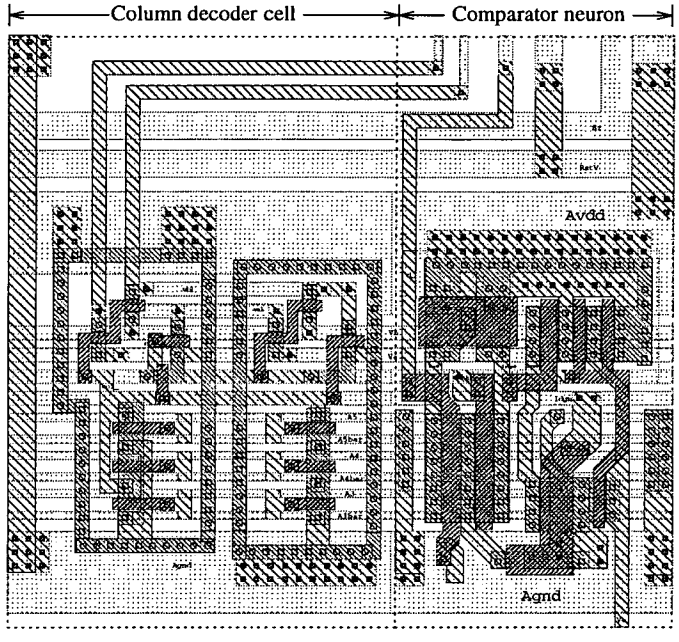


Figure 5.5 - Column decoder + neuron layout: 119μm x 106μm †

5.2.3. Row decoder with bank select

The row decoder is used during the two modes of chip operation:

(i) Programming: Selects a single row from the eight available. This is used in conjunction with the column addressing to select the device that receives the programming pulses.

(ii) Recall: Select a bank of four/six/eight rows to be used as inputs to the neuron at the foot of the column. For example, if the application only requires four inputs, then by selecting only the bottom four rows, full use is made of the neuron's input dynamic range. If all eight inputs were used, with four tied to zero volts, then the dynamic range would be halved, with eight integration capacitors connected together but only four active synapses.

The row decoder schematic is shown in figure 5.6.

† The comparator part of this cell was laid out for the 1.0μm process by Steve Churcher just prior to his departure from the Neural Network group.

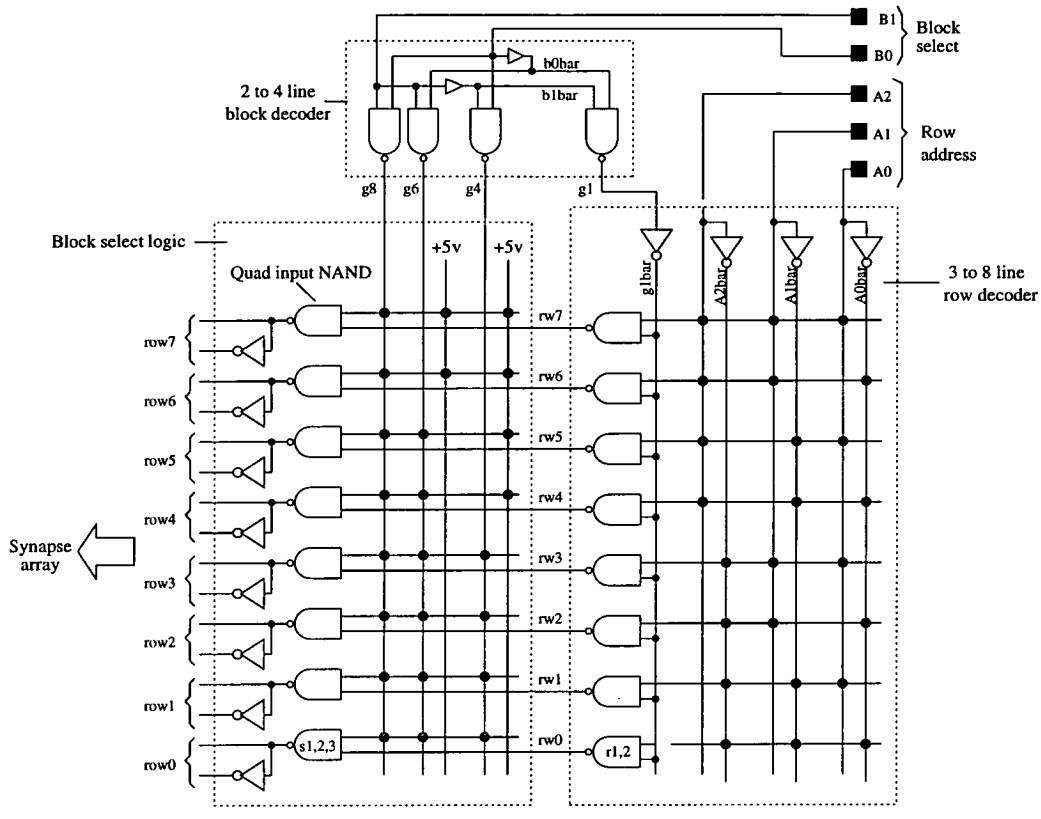


Figure 5.6 - Row decoder schematic

The two bit block select code (B1 B0) is used to set the number of input rows - four, six or eight.

The row decoder is built from eight instances of a common cell, with the functionality being determined by an additional instance that connects the gate inputs either to the address or to the address_bar line.

5.2.4. Synapse design

Before considering the ASiTEST3 synapse itself, there is one result from the ASiTEST2 chip that can be added to the a-Si:H design rules listed in chapter 3.

- The synapse cell in which the a-Si:H device was placed above the local integration capacitor works as well as the one in which it was placed over empty substrate; the former produces a more compact synapse cell.

The ASITEST3 synapse is based on the Schurch design used in ASiTEST2. However, the 1.5μm process used for the ASiTEST2 chip was no longer available, so all the cells had to be redesigned for the ES2 1.0μm process. It was decided to use this opportunity to improve some aspects of the ASiTEST2 synapse design.

• Divide by five: In the ASiTEST2 synapse, there was a need to scale the stored current by a factor of a sixth before it was used to generate the weight current, Iw. This was implemented as a 12/2 transistor mirroring to a 6/6 one. The effect of channel shortening,

produced by lateral diffusion of the source/drain implant, meant that the scaling achieved with this arrangement was actually nearer a tenth than a sixth. The scaling was also dependent on the magnitude of the current being scaled. In the ASiTEST3 synapse the scaling, this time by a fifth, is achieved with five 3/10 transistors mirroring to one 3/10. The results of a simulation comparing the performance of the ASiTEST2 and ASiTEST3 designs is shown in figure 5.7.

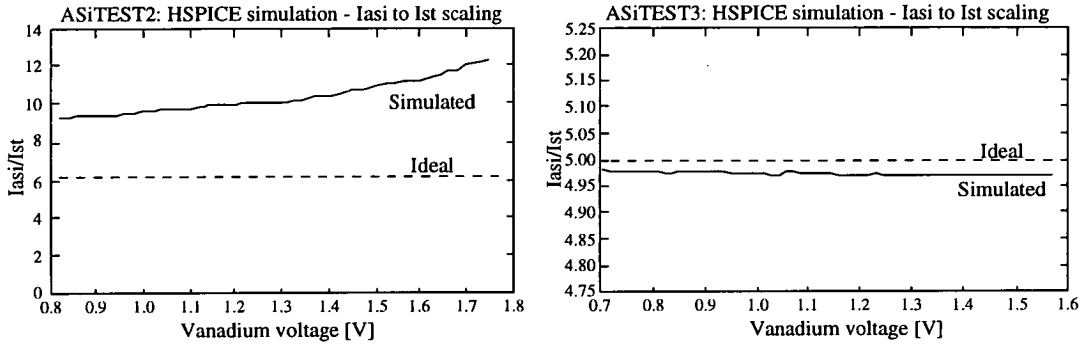


Figure 5.7 - Comparison of current scaling in ASiTEST2 and ASiTEST3 synapses

• Effect of activity voltage: In the Schurch synapse, the "sum" node voltage varies with the level on the integration capacitor. This produces changes in the drain source voltage across the transistors Msz and Mst, which in turn causes the mirrored currents Ist and Isz to be altered. On ASiTEST3, the transistors Mst and Msz were made long and thin to minimise this effect. The results of simulations comparing the performance of the ASiTEST2 and ASiTEST3 synapses at three different activity voltages are shown in figure 5.8.

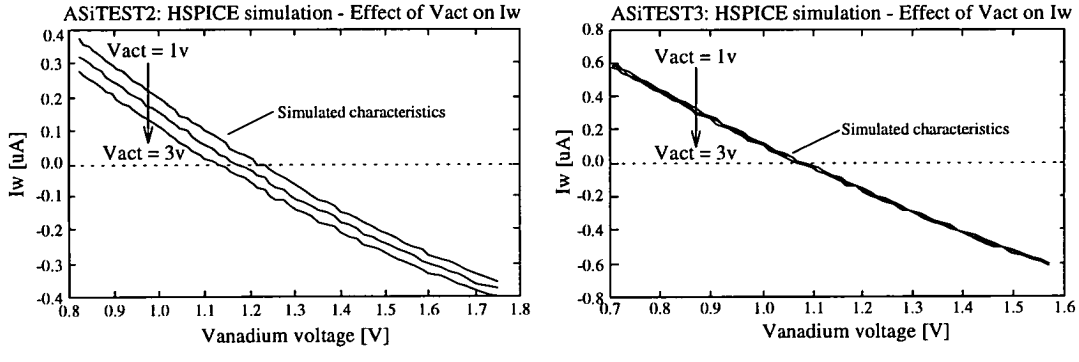
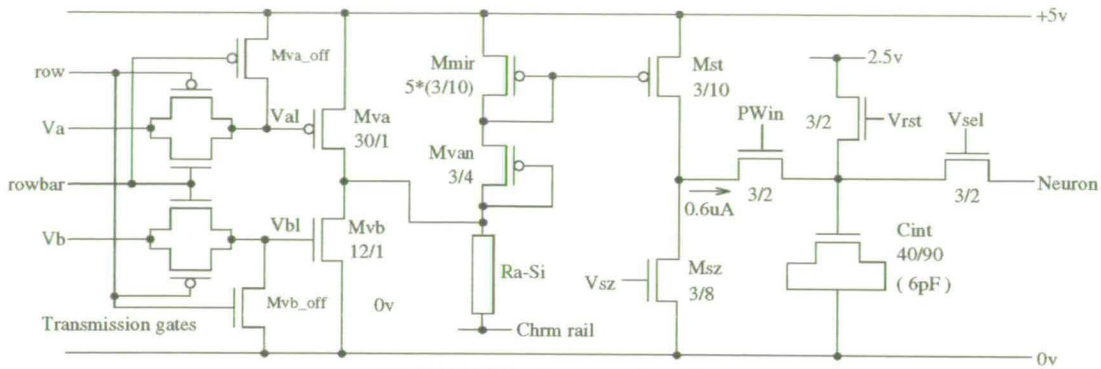


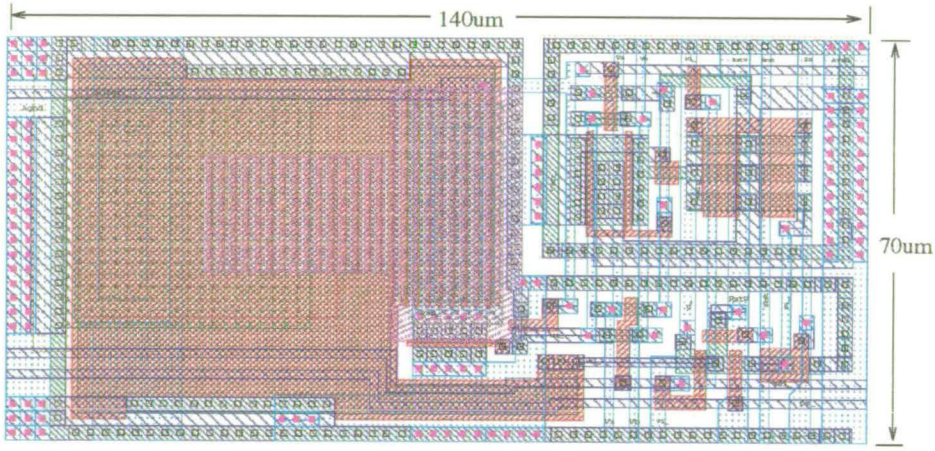
Figure 5.8 - Comparison of Iw currents in ASiTEST2 and ASiTEST3 synapses

• Layout: Before it was decided that the ASiTEST3 chip should consist of an 8 x 8 synapse array, the original intention had been to use a 16 x 8 synaptic array. To this end the synapse was laid out with a height-to-width ratio of 1:2, so that a 16 x 8 array would still from a square core region. When the decision was made to use an 8x8 array this synapse was re-arranged to give a roughly square shape, 105μm by 119μm. The original 16 x 8 synapse is much more compact and hence provides a better idea as to the minimum size of an a-Si:H pulsewidth synapse. Figure 5.9 shows the schematic of the

ASiTEST3 synapse and the 16 x 8 synapse layout.



(a) ASiTEST3 synapse - schematic



(b) ASiTEST3 synapse - 16x8 layout

Figure 5.9 - The ASiTEST3 synapse

The address transistors, Mva and Mvb, were chosen such that they supplied similar currents to the 1.5μm transistors used on ASiTEST2. As the figure also shows, the synapse cell now contains two transmission gates which only allow the Va and Vb gate voltages to be passed to the address transistors if the row select line is low. If the row select line is high then the voltages Va and Vb are set such that the programmer transistors are both off.

5.3. ASiTEST3 - Test board

A major part of the ASiTEST3 design was dedicated to the board used to test the 8x8 ANN chips. The board was designed such that it could take two ANN chips cascaded in series, whilst requiring only a minimum of external circuitry. The completed board needs: four external control signals; a 5 V power supply, taken from the PC; access to the PC's 16 bit data bus; programming pulses, again supplied by an HP pulse generator. The test board, illustrated in figure 5.10, uses static ram (SRAM) chips to store both the input/output pulsewidth signals and the ramp signal, Vramp.

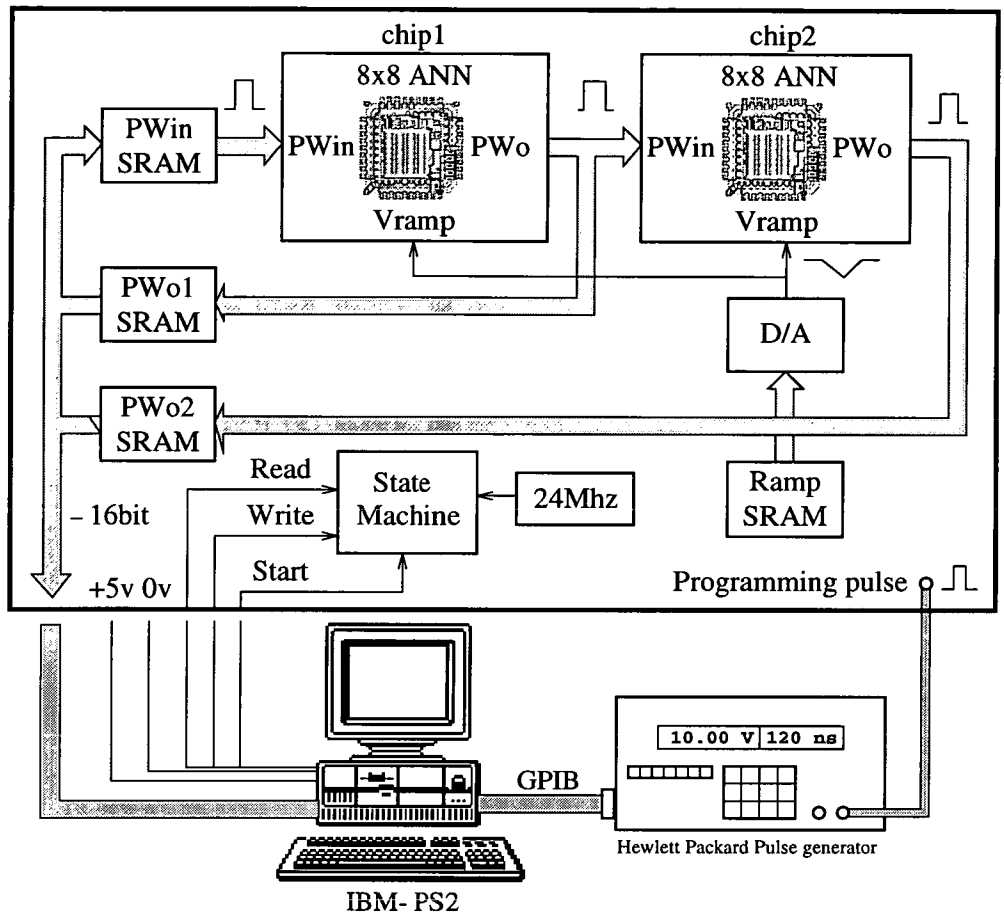


Figure 5.10 - ASiTEST3 test board block diagram

The test board contains a simple state machine which cycles through four different states. In state zero the PC can read/write to any of the SRAM chips on the board, so allowing inputs to be downloaded or results read back. After the Start control line is asserted, the board runs through the remaining three states independently. A typical sequence would be:

(i) State 0: Load pulsewidth input states into SRAM from PC.

(ii) State 1: Apply input pulsewidth signals to chip 1.

(iii) State 2: Use Vramp to generate threshold function for chip 1. The resulting pulsewidth output is applied as the input to chip 2, as well as being stored in SRAM.

(iv) State 3: Use Vramp to generate the threshold function for chip 2. Read resulting pulsewidth output into SRAM.

(v) State 0: Read pulsewidth output states from SRAM into PC.

In figure 5.11 the oscilloscope waveforms triggered by the Start signal are illustrated. The top trace shows an input pulsewidth. The middle trace shows the ramp signal which produces the output pulsewidth from chip 1, which is then used as the input to chip 2.

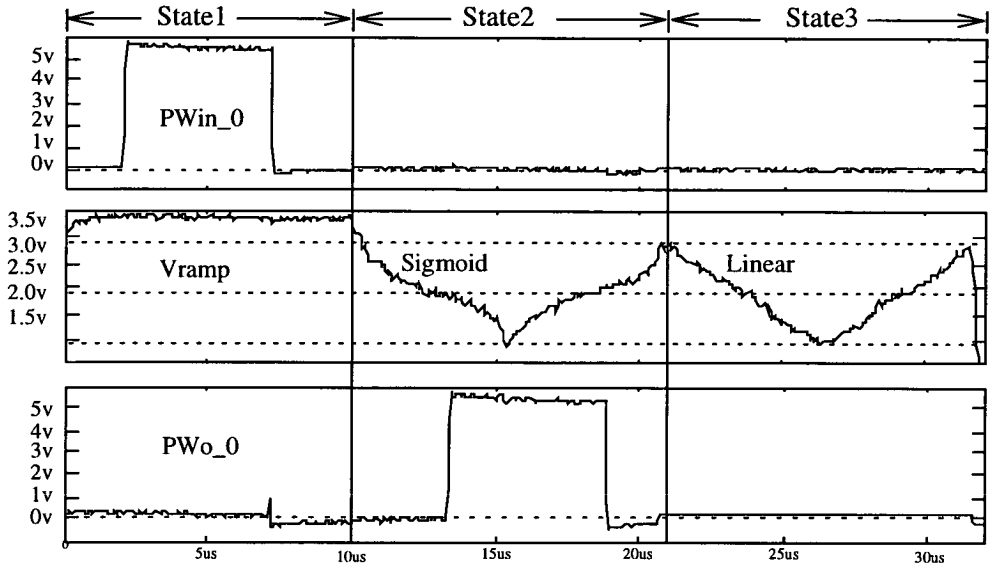


Figure 5.11 - ASiTEST3 test board scope waveforms

As figure 5.11 shows, different ramp signals, corresponding to different threshold functions, can be applied to chip 1 and chip 2 neurons. In the 32μs needed to cycle through states 1 to 3 the following operations are performed:

- Chip 1: 8 inputs * 8 synaptic weights = 64 multiply and add operations
- Chip 1: 8 activity voltages * sigmoid = 8 multiply operations
- Chip 2: 8 inputs * 8 synaptic weights = 64 multiply and add operations
- Chip 2: 8 activity voltages * sigmoid = 8 multiply operations

As this is a parallel architecture, the number of inputs and outputs could be increased without causing any increase in the cycle time.

5.4. ASiTEST3 - Results

The ASiTEST3 chips were supplied as four half wafers, which were cut into quarter wafers for processing as normal. As there were no other test structures, there was only one bonding configuration for this chip, detailed in Appendix C.

A total of four wafer segments were tested, each one corresponding to a different fabrication run. For each different segment a total of three devices were bonded up. The following table summarises the testing of these different wafers.

a-Si:H thickness	Date	Comment
820Å	1.8.94	Small number of devices formed: 10 out of 64 devices
945Å	26.8.94	Almost no forming: 2 out of 64 devices
645Å	26.8.94	Third of devices in array formed: 20 out of 64 devices
840Å	18.10.94	Almost all devices in array formed: 61 out of 64 devices

Table 5.2 - ASiTEST3 processing and test summary

The results from this chip are divided into four sections:

(i) Forming results

(ii) Switching experiments

(iii) Synapse multiplier characteristics

(iv) Complete ANN chip results

5.4.1. Forming results

The procedure adopted for forming was to select one of the devices using the address lines, and then to increase the height of the applied programming pulse until the device's resistance dropped to some lower value. However, it was found that devices other than the one addressed would sometimes form first. This is illustrated in the following set of results which show the resistance of each device in the 8 x 8 array after different forming pulses; the addressed cell is on column 0 row 0.

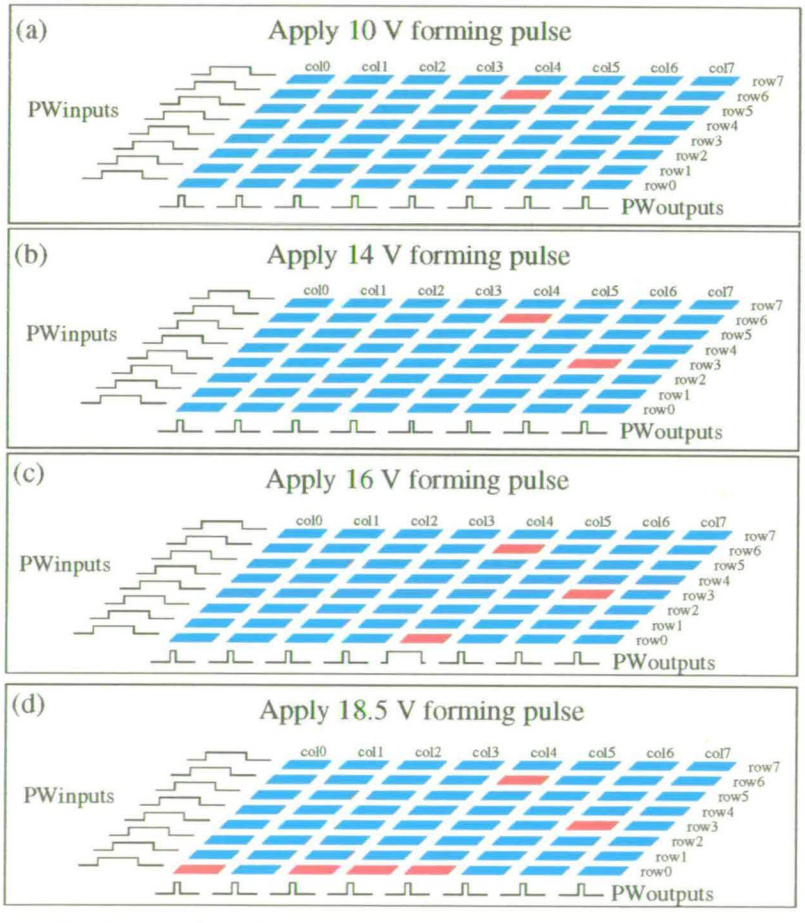


Figure 5.12 - 645Å chip, synapse array resistances during forming. Negative weights (high resistances) are blue (light grey) rectangles and positive weights (low resistances) are red (dark grey) rectangles.

The high voltages (> 16 V) required to form these devices exceeds the original specification of a 14 V maximum. At these high voltages the address transistors are effectively

bypassed, so the forming pulse appears across all the devices in the array.

The chips on which most devices could be formed all came from the 840Å wafer. The following set of results show the heights of the 300 ns forming pulses for each synapse cell. The light boxes represent devices that were not addressed when they formed.

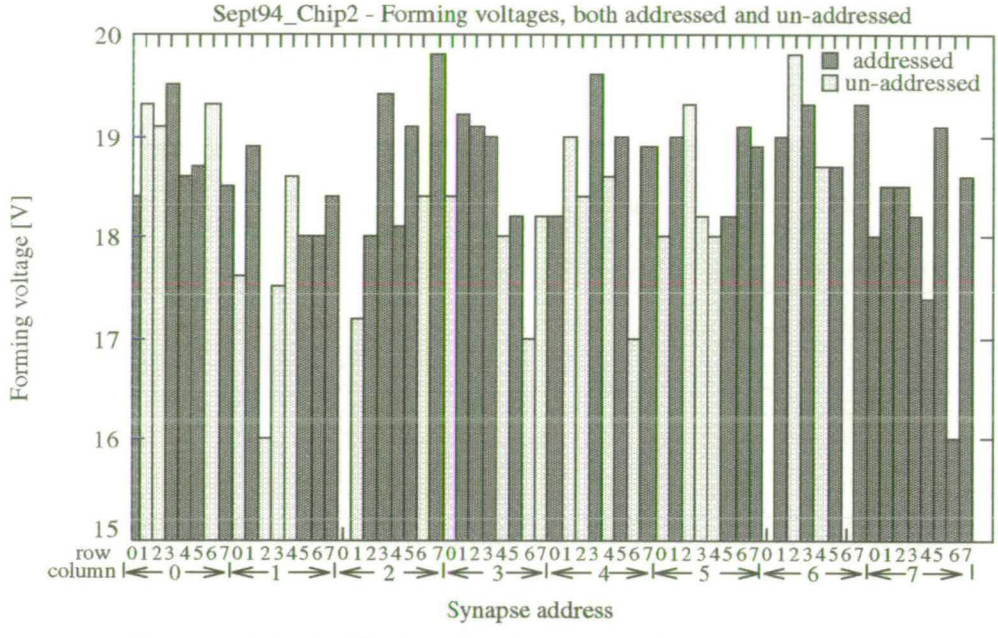


Figure 5.13 - 840Å chip distribution of forming voltages

There is no obvious pattern to identify the three devices which would not form on this chip. However, this was not the case on the 820Å and 645Å chips. On these chips there was always a complete row, or column, of formed devices, as the 645Å chip's synapse array, shown in figure 5.14, illustrates.

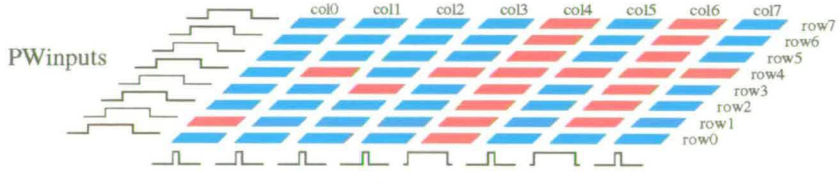


Figure 5.14 - 645Å chip 3 after forming.

The fact that the address transistors diodes are probably breaking down during the high forming pulses means that un-addressed devices are no longer isolated. This might mean that a column with more formed devices has higher currents, so favouring the forming of other devices in the same column.

It was also observed that a number of devices that formed into a low resistance state had returned to a high resistance state by the time the whole array had been tested. As some of these devices could not subsequently be switched, this implied that the high forming pulses, appearing across the whole array, had caused some of these low resistance, formed devices to go open circuit.

5.4.2. Switching experiments

In the ASiTEST3 system, pulsewidth outputs, rather than activity voltages, are used to monitor the state of the memory device in a particular synapse cell. As the pulsewidth output can range from 0μs to 10μs, a "zero" resistance, one which produces a current matching Isz, is indicated by an output pulsewidth of 5μs. A long pulsewidth, say, 10μs, indicates a low resistance device. A short pulse, say, 0μs, indicates a high resistance device.

The first two sets of results show typical resistance against pulse height graphs for the 945Å and 645Å devices. The results shown in figure 5.15 are from switching experiments carried out on one of the few 945Å synapses that actually formed.

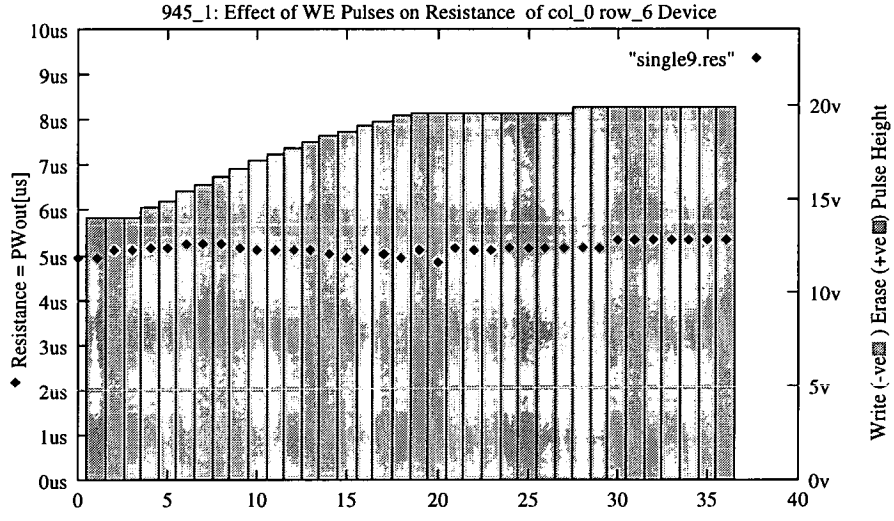


Figure 5.15 - Switching results from 945_1 chip.

As these results show, even though the device formed to a resistance state corresponding to 5μs, there is no further switching with pulses up to 20V.

The results shown in figure 5.16, are from a synapse on a 645Å chip.

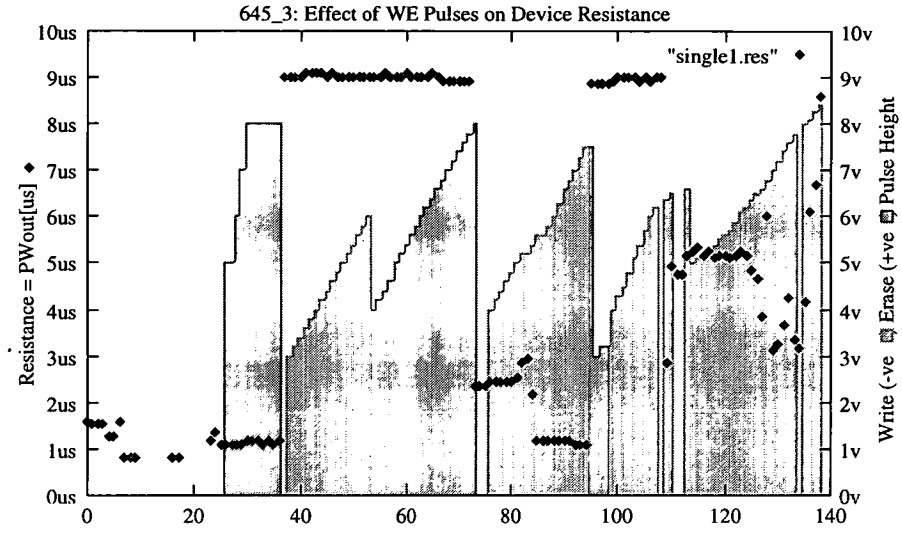


Figure 5.16 - Switching results from 645_1 chip.

In this case, the device switches from a maximum negative to a maximum positive weight with voltages in the range 5 V to 8 V. This was not the case for all the 645Å devices; some of them formed into a low resistance state and could not subsequently be switched.

The chips from the 840Å wafer were the only ones on which a useful number of resistors formed. On one of the three 840Å chips, forming was observed in a total of 61 out of the possible 64 synapse cells. The individual synapses on these chips were then characterised to test the range of resistances over which they could be switched. In the following set of results, the range of weight values over which each synapse could be switched repeatedly, say, 10 to 20 times, is shown.

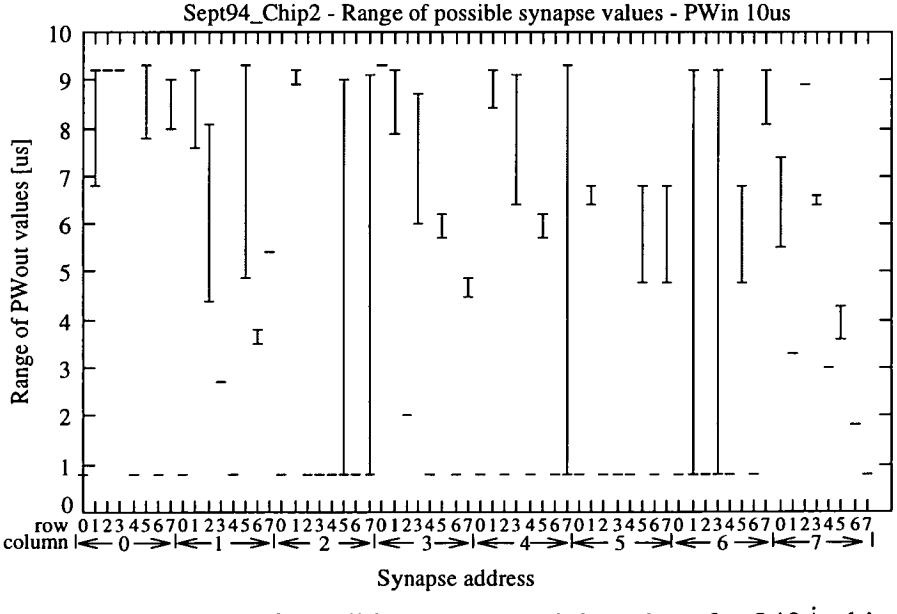


Figure 5.17 - Range of possible synapse weight values for 840Å chip 2

There were only five devices on the whole chip which could actually be repeatedly switched over the complete weight range. Many of the devices were restricted to a much

narrower range of weight values. The resistances corresponding to different programming pulses applied to one such device are shown in figure 5.18.

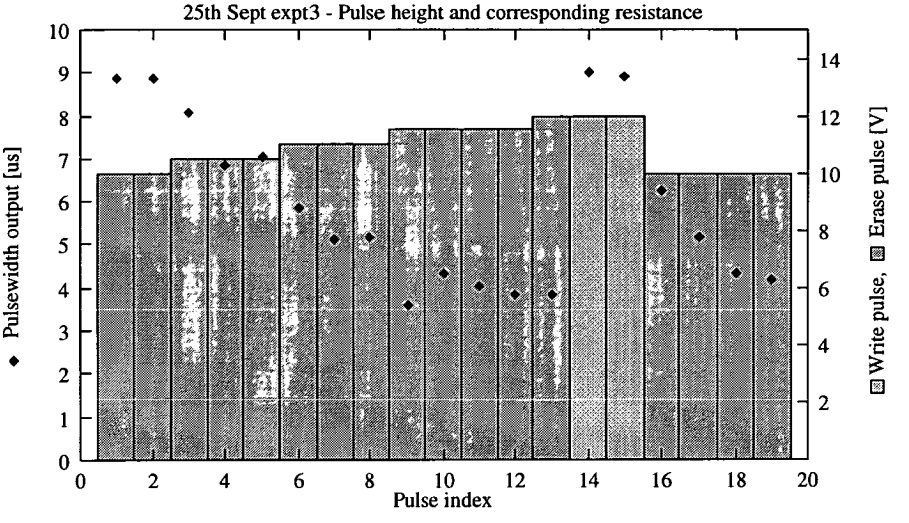


Figure 5.18 - Switching results from a 840A chip 3 synapse

During the initial erase pulses the weight gradually decreases (resistance increases) until the limit of its resistance range is reached. The device is then reset using a pulse of the opposite polarity and the exercise repeated.

During the discussion on the address circuitry design it was mentioned that transmission gates were being used so that analogue voltages could be passed to the address transistors. In the following switching experiment the programming pulse height remained constant while the voltage applied to the address transistor gate was gradually increased.

The address transistor gate voltages, Va and Vb, were controlled by an 8-bit digital to analogue (D to A) convertor, the output of which ranged from 0 V to 5 V. By varying the voltage supplied to the address transistors the maximum current that they can pass will alter. In the following set of results, the write/erase pulse was a constant +10 V. So, rather than displaying the height of the programming pulse, it is the 8-bit Vb value (0 to 255) that is displayed.

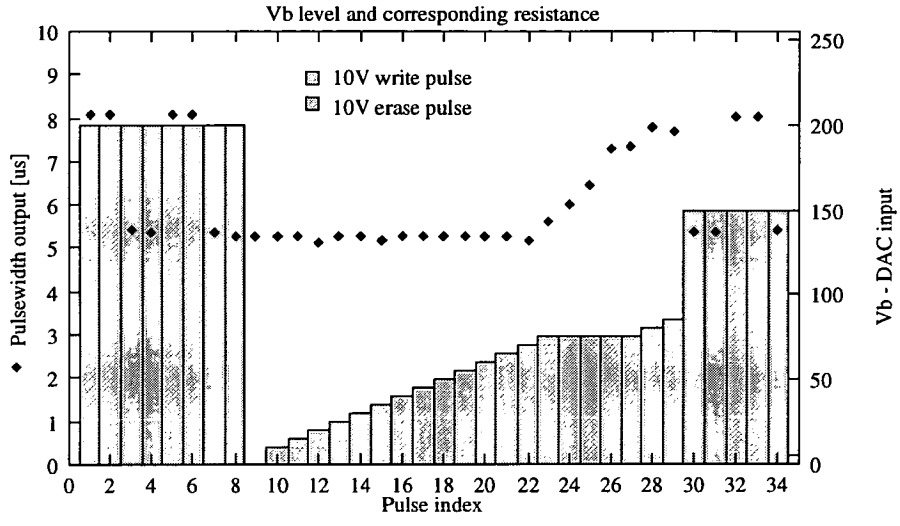


Figure 5.19 - Current based switching on 840Å chip

The first few pulses illustrate the effect of a high Vb level: the device toggled between two resistance states depending on whether it was a write or an erase pulse that was applied. After the eighth pulse the level of Vb was greatly reduced, with the 10 V write pulse now having no effect. The Vb level was then gradually increased, until the device switched into a new resistance state. With Vb at this level the device switched through a number of different states before reaching the limit of its range.

The reason for investigating this alternative programming method was, again, a system level consideration. If a complete system, including programming, is be integrated onto a single board, then programming circuitry would have to be designed to replace the HP pulse generator. It is simpler to construct a system in which an analogue voltage is used to drive transistor gates with the programming pulse of fixed amplitude, than it is to construct one with a variable amplitude, high current programming pulse. If it were possible to fabricate a-Si:H devices in which the switching voltages were as low a 5 V, as on the original glass substrate test chips, then the fixed amplitude programming pulse could be generated using standard digital logic. Indeed, such a programmer circuit was designed and constructed but was rendered unusable by the high switching voltages needed on all the test wafers available.

5.4.3. Synapse characteristics

As with ASiTEST2 the first non-switching experiment was a comparison of the synapse multiply characteristic obtained using a variable resistor with that produced by a synapse containing an a-Si:H device.

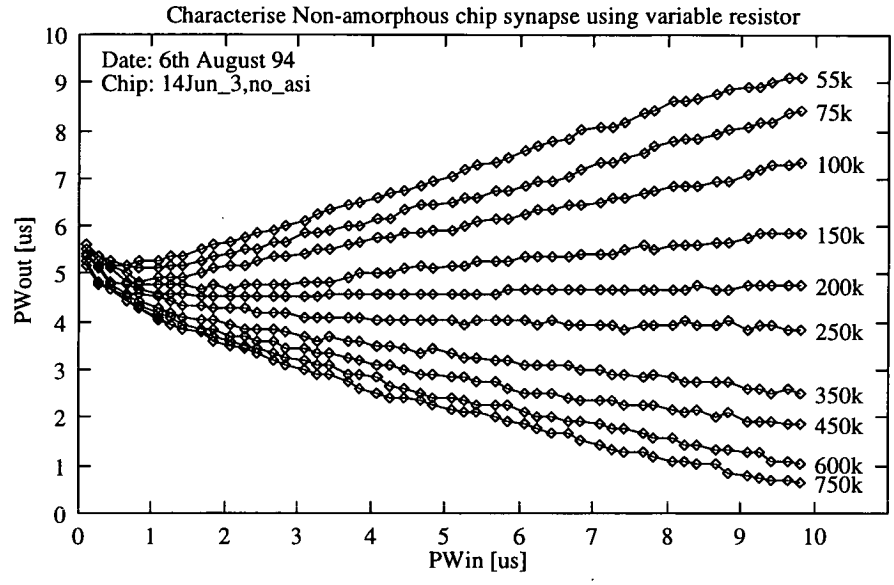


Figure 5.20 - Synapse multiply characteristic generated using a variable resistor

As figure 5.20 shows, the synapse covers quite a wide resistance range from 55 kΩ to 750 kΩ. The equivalent characteristic for a synapse with an a-Si:H resistor is shown in figure 5.21.

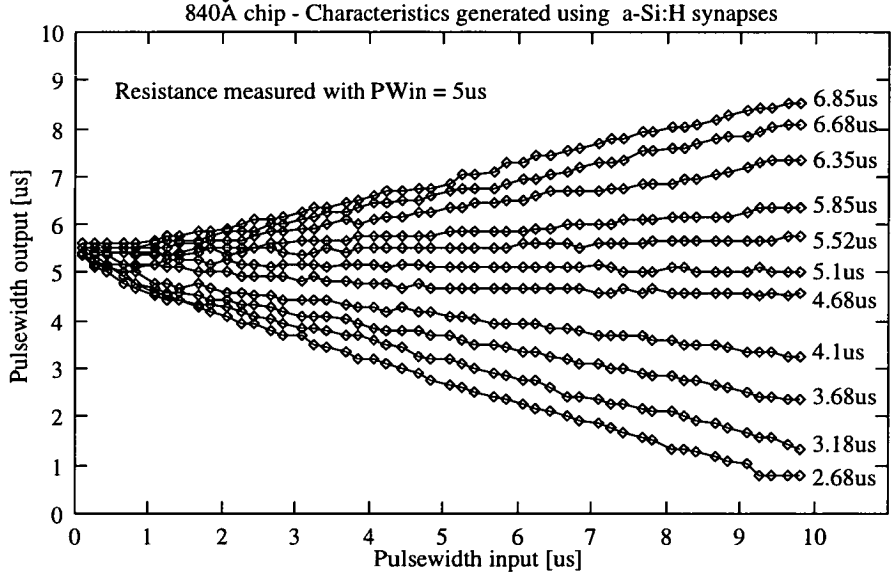


Figure 5.21 - 840Å - Characterisation using a-Si:H resistors

As with the ASiTEST2 synapses there is no major difference between the performance of the cell with the a-Si:H and the one without.

The synapse characteristic can also be used to compare the effect of the different Vramp threshold functions. Figure 5.22 shows three different weights characterised using a linear ramp and then using two sigmoid functions of different gain.

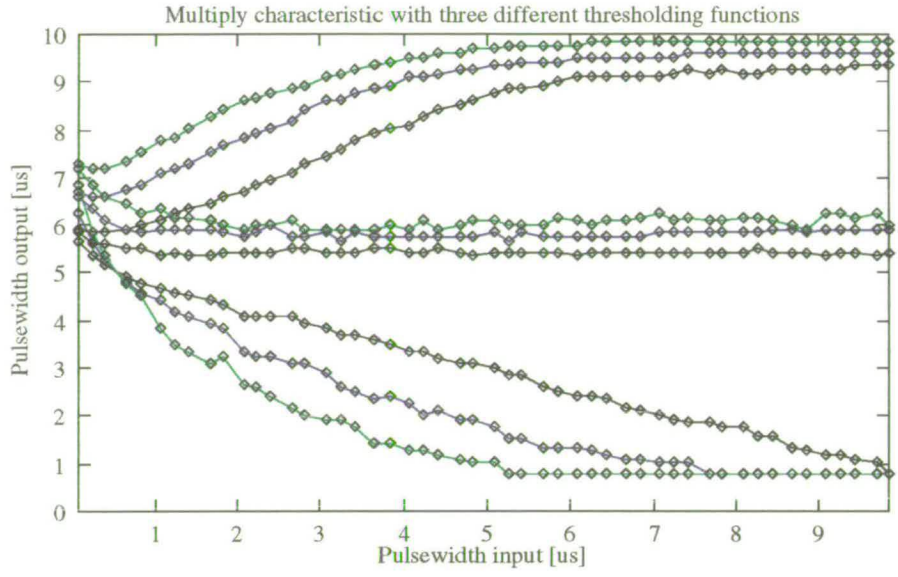


Figure 5.22 - Synapse characteristic generated using three thresholding functions

As one would expect the two sigmoid functions cause the output to saturate at a lower pulsewidth input than the linear ramp.

5.4.4. Complete ANN system

The ASiTEST3 board was designed to hold two ANN chips, cascaded in series. As there was no direct connection to the chip 2 inputs, a method for characterising the chip 2 synapse array had to be found. The solution was to set all the inputs to chip 1 to zero. The chip 1 activity capacitors then all remained at 2.5 V and all the chip 1 outputs were 5μs. This uniform set of signals was then used to characterise the chip 2 synapse array, as figure 5.23 shows.

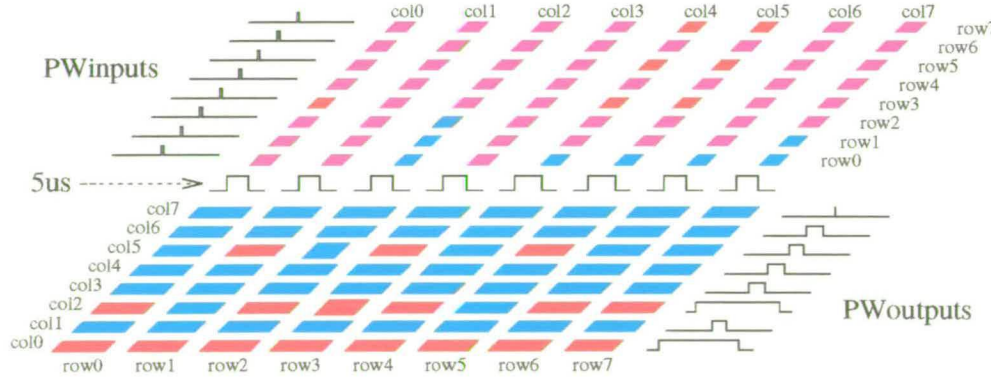


Figure 5.23 - Technique for characterising the synapse array on the second chip

The final experiment carried out on the ASiTEST2 chip was a comparison of the synaptic multiply characteristic for single and then double synapses. On the ASiTEST3 chip it was only possible to characterise: (i) single rows; (ii) the bottom four rows; (iii) the bottom six rows; (iv) all eight rows together. The following set of results show the output

pulsewidths generated when a synapse array was characterised using each of these different alternatives.

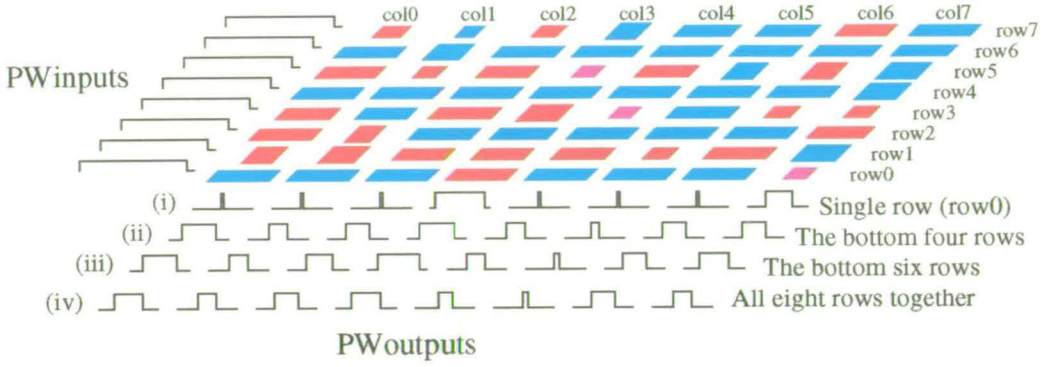
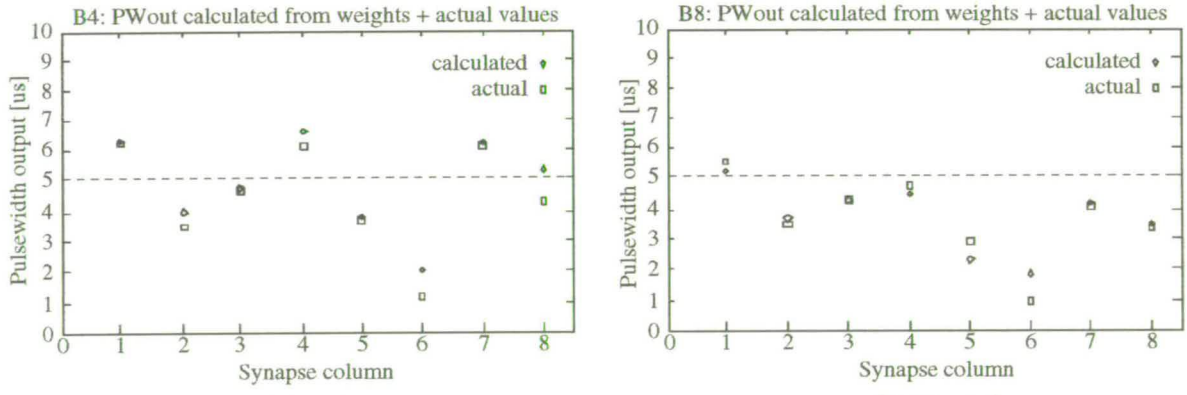


Figure 5.24 - Comparison of pulsewidth outputs for the four different input modes

The output pulses corresponding to the single row (row0) illustrate that negative weights produce narrow output pulses and positive weights produce wide output pulses. To see the effect of switching in more rows consider the output from column 7 with six and then eight inputs: as more negative weights are switched in, the output pulse becomes narrower.

The performance of the network during this experiment can be seen more clearly by plotting the expected output, calculated using the individual synaptic weights, with the performance when the rows are actually connected together.



(a) Block 4 (b) Block 8

Figure 5.25 - Comparison of expected output, calculated from weights, with actual output values

In general the calculated output is almost the same as the measured value. It is difficult to determine whether deviations are due to noise - these results are all single measurements with no averaging - or the effect of variation in the size of the integration capacitors: individual Iw values are measured using the capacitor local to a synapse cell, this may differ from the average capacitance once all the cells in a block are connected together.

The final test of the ANN functionality was to compare the output pulsewidths generated using different thresholding functions. The set of results shown in figure 5.26 were generated using first a linear ramp and then a sigmoid.

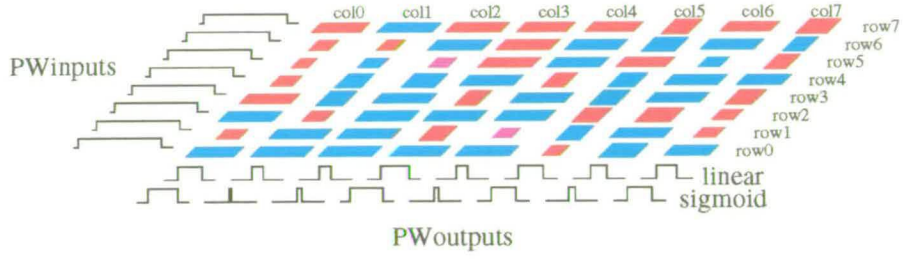


Figure 5.26 - Comparison of pulsewidth outputs generated using different threshold functions

The saturating effect of the sigmoid means that outputs that were just below 5μs with the linear ramp become much shorter and those that were above 5μs, such as column 3, become longer.

5.5. Discussion

The aim of the ASiTEST3 chip was to construct a small ANN which used a-Si:H memory devices for synaptic weight storage. The poor yield, in terms of synapses with a fully programmable weight range, meant that complete weight sets could not be downloaded. However, the arbitrary weights produced by the forming and switching experiments did allow various aspects of the chip's performance to be tested.

A possible explanation for the limited number of programmable devices was the high voltages required during forming. These high voltages probably appeared across already formed devices, causing them to go open-circuit. Two possible solutions to this problem would be:

(i) Moving to a process in which the transistors have a higher breakdown voltage. However, High Voltage CMOS (HVCMOS) processes are more expensive than the standard digital one used for the fabrication of the ASiTEST chips.

(ii) Producing a-Si:H devices that require lower forming and switching voltages. This would mean identifying the reasons for the high forming voltages, when compared with the original glass substrate devices. Possible areas for investigation might include: step coverage, parasitic resistances, differences in processing, the thermal conductivity of the substrate and the current limiting effect of the address transistors.

In terms of the system level aspects of the ASiTEST3 design the chip was extremely robust with no adjustment needed to any of the references after the initial setting. The linearity of the synaptic multiply was also quite surprising given the fact that the chip was

powered from a PC's 5 V rail which is extremely noisy. The simplicity of the test board interface made it very easy to characterise two whole chips, 128 a-Si:H devices, very quickly.

Chapter 6

Discussion and Conclusions

6.1. Introduction

The original aim of the project detailed in this thesis was to replace the capacitor in a dynamic storage synapse with an a-Si:H resistor for two reasons:

i) To produce an ANN chip with non-volatile storage

ii) To improve the understanding of the a-Si:H memory device by studying its performance in a practical application.

In this concluding chapter the progress made toward both these objectives is discussed, starting with results concerning the practical application of the a-Si:H memory device.

6.2. The a-Si:H memory device

Prior to the start of this project a-Si:H memory research focused on the switching behaviour of individual memory devices. This meant that during the course of the project a number of previously unexplored issues, related to programming and operating an array of memory devices, had to be considered. Although factors such as the effects of temperature, programming accuracy and long term stability are important in the development of practical a-Si:H devices, the issues that were raised during the design of the ASiTEST chips were:

1. Parasitic programming - where a non-addressed device receives a percentage of the programming pulse intended for other devices, potentially causing an unwanted resistance change.

2. Programming strategy - the approach used to decide the height of the programming pulse needed to program the device to a particular resistance.

3. Operating regimes - a set of conditions under which the memory device does not change resistance state. At the start of the project the sole operating regime was one where the applied voltage across the device was kept below 0.5 V.

4. Modelling switching behaviour - Necessary to simulate the memory device in conjunction with different address and synapse circuits.

5. Forming and switching yield - The number of devices in a formed array that can be programmed over the chosen weight range.

The following three sections contain a brief summary of the results from each of the three test chips associated with these issues.

6.2.1. ASiTEST1

The first test chip, ASiTEST1, was used to show that working a-Si:H memory devices could be fabricated on the surface of a CMOS chip; as well as a number of two-terminal test structures it also contained three designs of addresser cell, designed to investigate the problem of parasitic programming. During testing, however, it was found that even the a-Si:H devices in cells with the address transistors turned on could not be programmed. In order to try and understand this inability to program the a-Si:H devices associated with address transistors the device's switching behaviour was then investigated using the two terminal test structures.

In the course of these switching experiments it was observed that a large current pulse accompanied each resistance change. The lack of switching in cells with address transistors could then be explained by the current limiting effect of the MOSFETs. This result pointed to a solution to the problem of parasitic programming: only if the address transistor in a cell was on, and if it could supply enough current, would the device change state. Larger address transistors were used on the ASiTEST2 and ASiTEST3 chips which allowed the associated a-Si:H devices to be programmed.

While a resistance change was always accompanied by a large current pulse there were also occasions when there was a current pulse but no resistance change. This characteristic, that of a threshold switch, where the on-state is only sustained while the bias is above a critical holding point, was catered for by the model developed to simulate switching behaviour. This model was used to simulate the a-Si:H device's switching behaviour during the design of the ASiTEST2 and ASiTEST3 chips.

During the switching experiments it was also observed that the device's switching behaviour was not as predictable as that reported for the original "glass substrate" devices, where the final resistance was a function of the applied pulse height. On the ASiTEST1 chip, a pulse that had earlier caused a minimal resistance change could next time cause the device to switch from the lowest to the highest resistance state. The programming strategy adopted for the ASiTEST chips was therefore one where the height of the programming pulse was increased gradually until a resistance change occurred. This iterative programming was obviously much slower than that that could be achieved using the original one-shot approach. It is possible that the change in the performance of the memory device was due to factors other the change of substrate: during the course of the project changes were made to both the deposition equipment and the material used for the top vanadium contact.

While characterising the two terminal devices it was noticed that they were stable, that is their resistance remained unchanged, during current sweeps which did not exceed 50 μA. During these sweeps the voltage across the device rose to 5 V, much higher than the original 0.5 V operating limit. A new "current-limited" operating regime was therefore introduced, in which it was the current through the device that was restricted rather than the voltage across it.

The results from the ASiTEST1 chip were used as the basis for the design of the a-Si:H based synapse circuits contained on the ASiTEST2 chip.

6.2.2. ASiTEST2

The ASiTEST2 chip was used to test five different designs of a-Si:H pulsewidth synapse: three based on the EPSILON synapse and two on a distributed capacitance design. The need to submit the final design, ASiTEST3, meant that the testing done on the ASiTEST2 chip was the minimum required to choose the most suitable synapse for this final chip. Nevertheless, a number of results concerning the memory device were still obtained during the course of this testing.

In both types of synapse cell the a-Si:H resistor could be switched into different states that spanned the whole range of synapse weights. Once a device, now operating in the current regime rather than the original 0.5 V voltage regime, had been programmed it was quite stable.

The programming voltages needed to change the device resistance were in the range 6 V to 12 V, much higher than the 2 V to 5 V required for the two terminal devices. The need for higher programming voltages was due partly to the voltage dropped across the address transistors in series with the device.

The synapse multiply characteristics of the different designs were all very similar. However, the EPSILON cells were more difficult to set up and required more support in the way of control signals and power supplies. For these reasons it was decided that the final chip should be based on the distributed capacitance synapse rather than the EPSILON synapse, as was the original intent.

6.2.3. ASiTEST3

The ASiTEST3 chip was designed as a complete ANN based on an 8 x 8 array of synapses. It was designed with ease of testing in mind and hence had fully digital addressing and a minimum of control signals.

Chips from four different a-Si:H fabrication runs were tested. The last run, where the a-Si:H thickness was 840Å, was the most successful. On one of the 840Å chips 61 out of the 64 devices in the array formed. Unfortunately, only five of these could then be

switched over the whole weight range. This poor yield of switchable devices was attributed to the high programming voltages which were in excess of the address transistor diodes breakdown voltage. It was therefore likely that high forming voltages were appearing across already formed devices causing them to go open circuit. This problem could most easily be overcome by using a high voltage CMOS process in the construction of the backplane. The chip itself functioned satisfactorily and was very simple to setup and test.

6.2.4. Results summary

The results from the different test chips can best be summarised in terms of the application issues discussed in section 6.2:

1. Parasitic programming - As switching requires a combination of high voltages and currents non-addressed devices in array will not be reprogrammed.

2. Programming strategy - The strategy adopted was to gradually increase the height of the programming pulse until a resistance change occurred. This iterative process somewhat negates the advantages offered by the device's fast programming time.

3. Operating regimes - Two operation regimes were used, one defined by a current limit and one by a voltage limit.

4. Modelling switching - A model was developed that allowed the device to be simulated in conjunction with CMOS circuitry.

5. Forming and switching yield - While the majority of the devices in an array could be formed only a small number of devices could be programmed over the complete weight range.

6.3. Non-volatile synaptic weight storage in ANNs

The primary justification for this research was to produce an ANN chip with a-Si:H based non-volatile storage. While such a chip was indeed developed, results from the project suggest that the a-Si:H memory might be more suited to a more tightly focused application.

In order to provide a framework for the following discussion on suitable memory technologies for ANN chips the intended application is considered. One of the early justifications for producing analogue neural chips was to produce hardware accelerators for implementing algorithms already running on conventional digital computers. However, experience with large analogue neural chips, such as EPSILON and Intel's ETANN, suggest that the problem of interfacing such a chip to the host computer means that it is easier to use dedicated digital hardware for such a task. However, there other applications in which analogue hardware still offers advantages.

- Small, robust chips for tasks such as remote monitoring. For example, the Kakadu chip used for monitoring heart arrhythmias[3].
- Large, dense networks for tasks such as template matching. For example, the Synaptics chip used for reading cheque codes[81].

Both these types of chip, along with suitable memory technologies, will now be considered in turn with a view to selecting the one for which the a-Si:H memory is most suited.

6.3.1. Designing a small, robust ANN chip

The "small and robust" ANN chip could best be thought of as an intelligent A to D convertor: the chip has analogue inputs, which are processed by the neural network, and produces its output in a form that can be read by a digital controller. As the design specification for such a chip would be similar to those used in the design of the ASiTEST3 chip this design will be used to illustrate two possible application scenarios.

The first application is as the data gathering element in a remote sensing situation, as illustrated in figure 6.1.

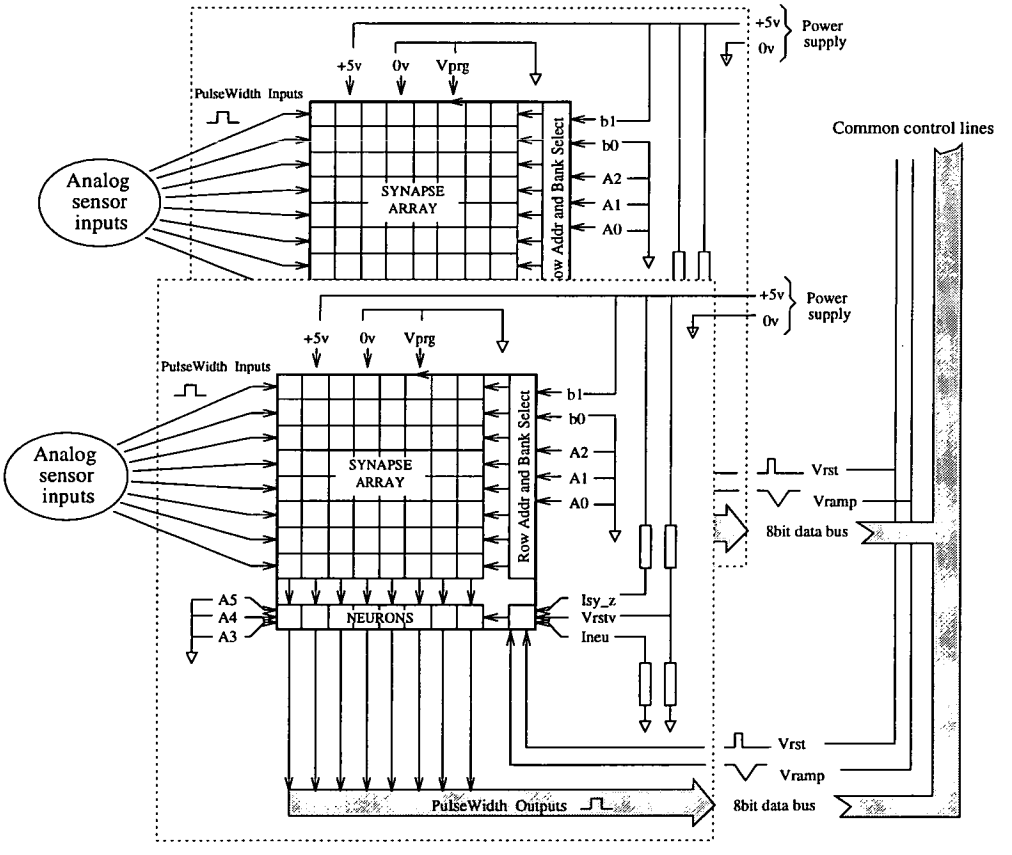


Figure 6.1 - Remote sensing application

As figure 6.1 shows, the only local support circuitry required by such a chip is 2 to 3 fixed value resistors, used to set reference currents. By comparison a dynamic storage ANN would require external refresh circuitry and a separate EEPROM to hold the

synaptic weights. The use of pulsewidth outputs means that results could be read back by the controlling computer using extremely simple digital circuitry[82].

Another possible application for this "small and robust" chip might be in an autonomous robot. As figure 6.2 shows, a single ASiTEST3 chip could be used to interface directly between analogue sensors and the robot's motors. The only additional support circuitry that is required is a small timer circuit to generate the reset and Vramp waveforms. The robot's response to input stimuli would then be determined by the weights stored in the synapse array.

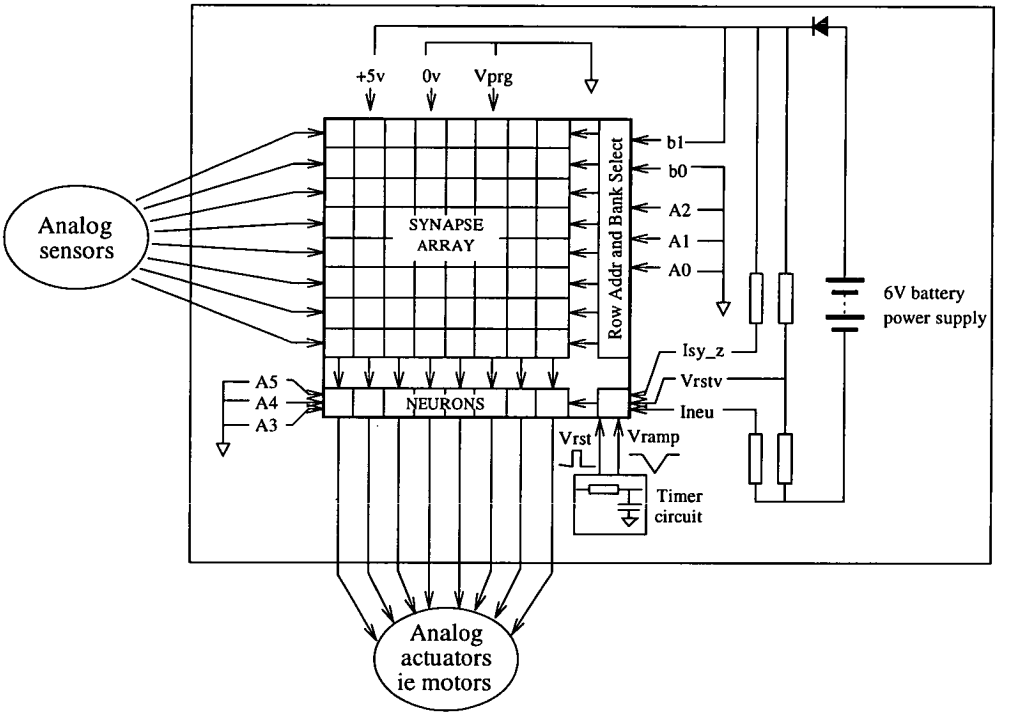


Figure 6.2 - Possible autonomous application

Having considered possible application areas for such a chip it is now necessary to decide on the most suitable memory technology for the synapse array. The specification for the memory technology would have to include low cost, reliability and 5 V operation. The two most suitable memory technologies would seem to be battery backed SRAM and floating gate EEPROM.

Battery backed SRAM offers the flexibility of dynamic weight storage: entire weight sets can be downloaded digitally, with low power operation. In addition modern batteries have a lifetime of up to five years.

If a truly non-volatile technology is needed then floating gate EEPROM would seem to be the obvious choice:

- Four quadrant multiplication can be performed using a two transistor synapse by taking advantage of the transistor's Vds/Ids characteristics.

- Floating gate cells can now be constructed using a standard CMOS process so no specialised processing is needed.
- The fact that it is based on a proven digital technology means that the effects of temperature, cycling and hold time are well understood.
- The ability to make small incremental changes to the charge stored on the floating gate make it suitable for on-chip learning.
- The low currents during programming make the integration of on-chip programmer circuitry feasible so allowing 5 V only parts.

This use of EEPROM cells in standard CMOS to produce low cost neural chips has been the approach adopted by the American company Synaptics which specialises in neural VLSI[81]. If the a-Si:H resistor is considered using the same criterion then it is obviously not the best choice for small, CMOS based chips:

- The memory cell is not compact as large address transistors are required and the device must be constructed on a "flat" surface.
- The a-Si:H processing, although simple, is specialised and certainly different from the standard CMOS furnace oxide cycle.
- The switching mechanism is not understood and neither are the effects of temperature, repeated cycling or long term retention.
- Switching is occasionally erratic making it unsuitable for on-chip learning
- The high programming voltages mean that either a high voltage CMOS or bipolar backplane would have to be used in order to isolate devices in the array during programming. The high programming currents would also make it difficult to integrate the programmer circuitry on chip giving 5 V only parts.

Now consider the "large and dense" neural chip which is more suited to the a-Si:H resistor technology.

6.3.2. Designing a large, dense ANN chip

To-date the ANN chips with densest synapse arrays have been binary fixed weight network that were either mask or one-time programmable. By using the a-Si:H memory it would be possible to construct a high density chip with analogue, reprogrammable weights. Consider the advantages of the a-Si:H memory device:

- The active area of the device is extremely small, 1μm diameter. This is extremely compact, especially compared with the planar layout required to construct an EEPROM cell in thin-film technology[83].
- The programming pulses are very fast (120 ns).

- It is a thin film technology with the possibility of a single substrate containing more than one layer of memory devices.
- It is compatible with amorphous silicon photoresistors which can be used to generate an optical input.

If the a-Si:H resistor were to be used in such a chip then the synapse design would have to differ considerably from that used on ASiTEST3. Firstly, the synapse would have to be designed to be "fail safe" - in the ASiTEST3 chip a high resistance (open-circuit) is equivalent to a large negative weight. In the architecture shown in figure 6.3 the synapse cell only contains an a-Si:H resistor. Unprogrammed and open-circuit resistors thus both act as a zero weight "fail-safe".

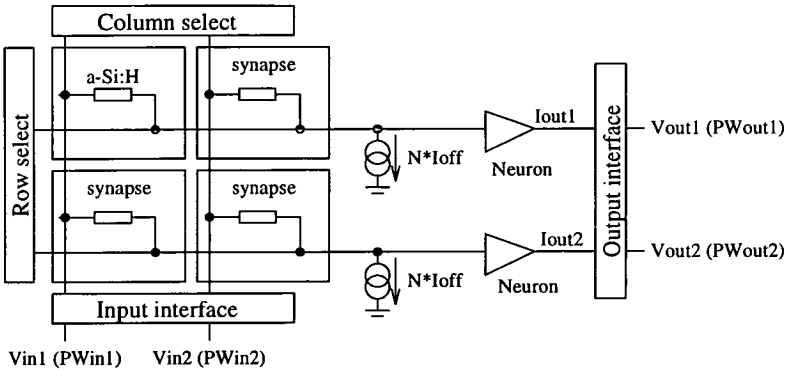


Figure 6.3 - Synapse for large networks

The area of this synapse cell would be defined by the crossing of the row column lines as in the fixed weight arrays. As the synapse array only has to connect to the address circuitry at the chip perimeter the vanadium and chromium rails could be used for row/column tracks. This would eliminate the large passivation openings used to connect the memory device to the synapse on the ASiTEST chips.

The use of addressing circuitry round the perimeter of the array, as shown in figure 6.3, would not be possible using a standard 5 V CMOS process as the high programming voltages would reverse bias any S/D diodes. A high voltage CMOS or TFT backplane would be required to cater for the high voltages and currents needed during programming. A TFT backplane would allow much larger arrays to be constructed than would be possible with crystalline technology.

In the proposed architecture inhibitory weights are achieved using a common current sink connected to each synapse column. This means that all the limitations of the BT network could be overcome (positive weights, external addressing, external op-amps). As a-Si:H is photo-conductive it would also be possible to build chips with optical input that did not require the use of external masks for weight definition.

These factors would all tend to suggest that this technology is more suitable for the large arrays currently associated with fixed weight networks rather than flexible CMOS based

designs.

6.4. Final conclusions

The a-Si:H memory device does not appear to be the most suitable memory technology for small CMOS ANN chips - as the original project direction implied. However, the potentially small synapse size and fast programming would seem to make it a suitable technology for large, dense networks, perhaps constructed using thin film drivers.

Appendix A

Analogue Storage using Floating Gate Technology

Introduction

While this thesis is primarily concerned with the use of a-Si:H analogue memory devices for synaptic weight storage, there are other areas in which non-volatile, analogue storage devices can be employed. Of the various technologies considered in chapter 2, the only one being used in commercial analogue chip designs is floating gate EEPROM.

Many analogue chips now contain EEPROM cells although mostly for digital, rather than analogue, storage. For example, the 12-bit digital to analogue convertors (DACs), that use EEPROM cells in place of laser trimming, use the memory cells to hold a digital correction code, rather than an analogue voltage. There are, however, a few applications in which EEPROMs are used as analogue memory devices:

- Analogue storage of speech signals[84, 85].
- Offset compensation in op-amps[86].
- Synaptic weight storage[53, 87]

One obstacle to a more widespread use of EEPROM cells has been the need for specialised fabrication steps, required to grow the thin tunnel oxide used during programming. A number of recent publications have reported techniques for constructing EEPROM cells using a standard CMOS process. These new memory structures have been applied to various areas:

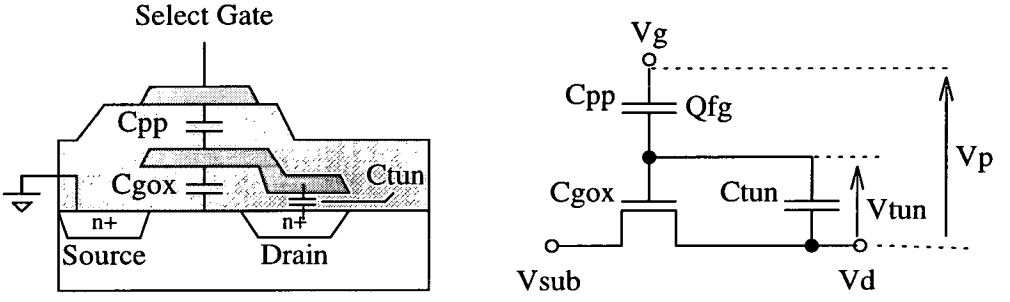
- The removal of fixed pattern noise in CMOS imager circuits[88].
- Offset compensation in op-amps[89, 90].
- Synaptic weight storage[91].

This appendix divides into two. The first section considers the programming of analogue EEPROM cells, in order to better understand what it is that makes this technology so appealing as an analogue memory. The second section contains a review of different floating gate cells that have been constructed using standard CMOS processes.

Programming EEPROM analogue memories

This section is concerned with the programming of analogue EEPROM cells. The discussion is divided into two areas: firstly, the geometric factors to be considered in the design of an EEPROM cell and secondly, the parameters of the programming pulse itself.

• Cell design[92] - An EEPROM cell is programmed using Fowler-Nordheim tunnelling through a thin tunnel-oxide region. The tunnelling current is dependent on three factors: the area of the tunnelling oxide region, A_{tun}, the thickness of the tunnelling oxide, X_{tun}, and the voltage across the tunnel oxide, V_{tun}.



(a) Cross-section (b) Schematic representation

Figure A1 - Representation of a EEPROM cell highlighting major capacitances

If a programming pulse of amplitude Vp is applied to the select gate of the EEPROM cell, then the tunnelling voltage V_{tun} will depend both on the capacitive coupling between Cpp, Cgox and Ctun, and the charge already stored on the floating gate.

During a write pulse:

$$V_{tun} = V_g K_w + \frac{Q_{fg}}{Ctot} \quad \text{where} \quad Kw = \frac{Cpp}{Ctot} \qquad \text{A1.1}$$

During an erase pulse:

$$V_{tun} = V_d K_e - \frac{Q_{fg}}{Ctot} \quad \text{where} \quad Ke = \frac{Cpp + Cgox}{Ctot} = 1 - \frac{Ctun}{Ctot} \qquad \text{A1.2}$$

and where Ctot is the sum of all the capacitances i.e. $Ctot = C_{pp} + C_{gox} + C_{tun}$

The tunnelling voltage can therefore be tailored by changing the coupling coefficients, Ke and Kw: if the capacitance Cpp is increased then Kw will increase resulting in a higher effective tunnelling voltage during write operations.

• The programming pulse - The effect of different programming pulses on the final threshold voltage of an EEPROM device can be seen by considering the results of two simulations, which were based on the EEPROM modelling equations derived by Kolodny[93].

In the first set of characteristics, figure A2, the effect of varying the width of pulses of different height is illustrated.

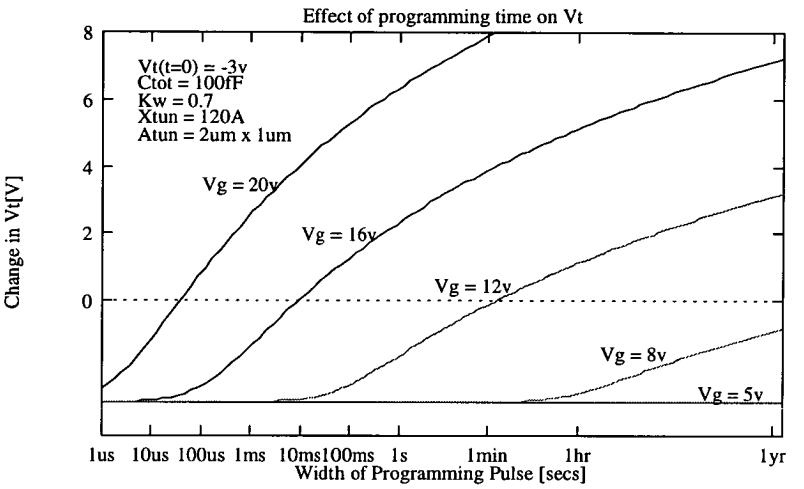


Figure A2 - Effect of pulse width on Vt for different values of Vg

As figure A2 shows, a standard EEPROM cell (X_{tun} = 120Å) can be programmed with voltages as low as 12 V, if pulses of the order of a minutes duration are used.

In the second set of results, figure A3, the pulse width is kept constant while the height of the applied pulse increases from 14 V to 20 V.

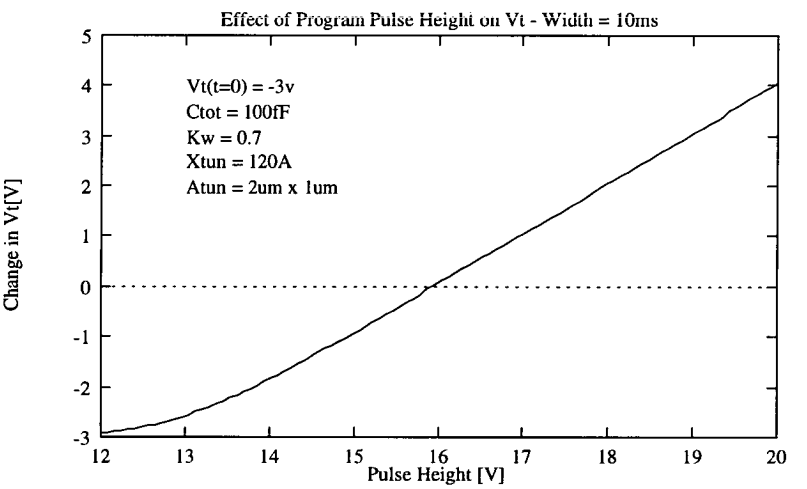


Figure A3 - Effect of pulse height on Vt

As figure A3 shows, there is an almost linear relationship between the pulse height and the change in the EEPROM cell's threshold voltage. By using this characteristic of EEPROM programming Hu[94] demonstrated that it was possible to use a commercial digital EEPROM for analogue storage. An analogue waveform was sampled at 10 ms intervals and the resulting voltage, scaled between 14 V and 20 V, used as the programming pulse for an individual bit of a 2K by 8-bit EEPROM.

Using a similar approach, but with a customised fabrication process, the US company Information Storage Devices[85] have produced speech storage chips based on analogue

EEPROM technology. The ISD1016 has 128,000 analogue cells and stores 16 seconds of speech at sample rate of 8 kHzs; equivalent to programming a 1000 x 128 synapse array in 16 seconds. The devices are specified for 10,000 read/write cycles and the chip requires a single 5 V power supply.

In applications where higher resolution, say 8-bits, is required the variation in the characteristics of devices across a chip, and the effect of repeated of cycling, make an it necessary to use an alternative programming strategy. By using an iterative programming scheme, with 50 ns programming pulses, Sin[84] has managed to achieve 8-bit resolution in less than 20 μs.

A number of specialised fabrication procedures have been suggested in order to improve the analog performance of floating gate devices. These include:

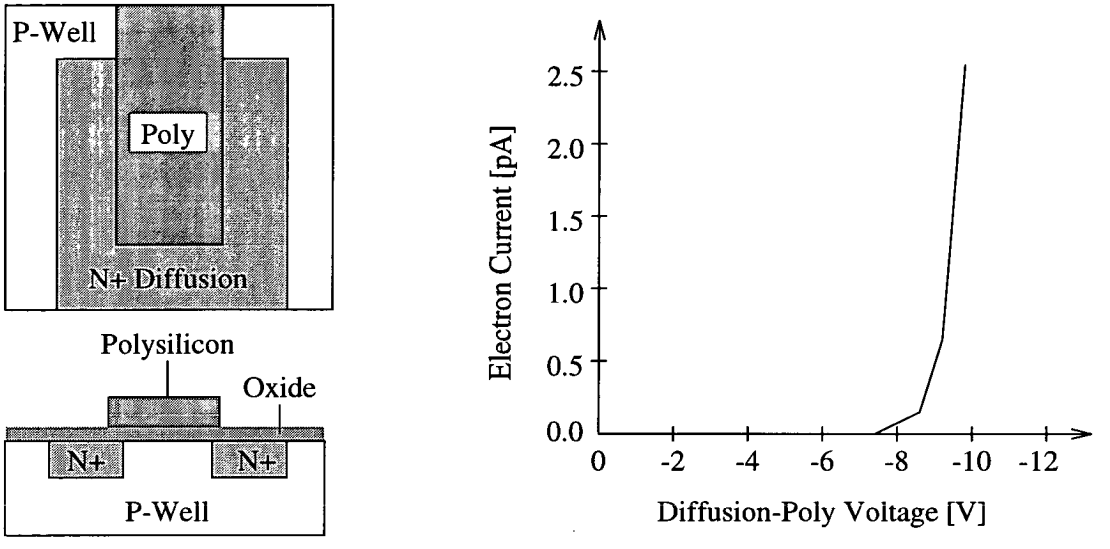
- A buried injector in conjunction with a sinusoidal programming scheme, suggested by Vittoz[6].
- A buried injector structure that allows 5 V programming, the VIPMOS cell[95].
- An ultra thin oxide with a control gate to achieve hot-electron programming, which is more linear than programming based on tunnelling[96].
- The inclusion of high value resistors to link a small injection capacitor with the main floating gate, allowing more accurate programming[97].

The disadvantage of all these new techniques is that they require specialised fabrication procedures, such as buried injectors and ultra-thin oxides. In the following section the design of floating gate cells using a standard CMOS process is considered.

Floating Gate in standard CMOS

Recently there have been a number of publications in which floating gate MOSFET cells constructed using a standard CMOS process have been described. A standard CMOS process is much cheaper than a conventional EEPROM one, and is more generally available.

The first standard CMOS floating gate cell was suggested by Carley in 1989[90]. He described a method for tunnelling through standard gate oxide at relatively low voltages: rather than using specially textured polysilicon to enhance the electric field, as discussed in chapter 2, he instead relied on geometric factors, determined by the mask layout. A test cell, referred to as a current injector, was fabricated using a 2μm p-well CMOS process. The thickness of the gate oxide for the process was 400Å. The current injector, shown in figure A4, consists of a polysilicon rectangle that ends in the middle of an area of n+ diffusion.



(a) Cross section (b) Programming characteristic

Figure A4 - Carley's current injector device

The electric field in the current injector is concentrated at the corners of the polysilicon rectangle, resulting in a field enhancement factor of between 2 and 4. This field enhancement allows tunnelling currents to be established with voltages in the range 8 V to 12 V; without field enhancement a gate/diffusion voltage of around 25 V would be needed to produce tunnelling through a 400Å oxide.

The tunnelling currents in Carley's cell are considerably smaller than those in a conventional EEPROM cell. The tunnelling current in Carley's cell was of the order of 1 fA; with a typical capacitive load of 250 fF, this small current results in a threshold voltage change of only 4 mV/s. One advantage that the standard CMOS cell does have over conventional EEPROM is that the thicker tunnel oxide has much better charge retention performance: Carley calculates a loss of only 0.1 percent in 10 years at an operating temperature of 100°C.

Carley's application of this analogue memory was as a trimming element in an op-amp circuit: using one of these devices the op-amp input offset voltage was trimmed from 10mV to 0.5mV. Bibyk and Ismail have also considered such memory devices, based on tunnelling through gate oxide, in the context of synaptic weight storage[98].

One disadvantage of using gate oxide as the tunnelling medium is that the write and erase characteristics are very different. In Carley's cell an erase voltage of 24 V is required before tunnelling can occur, compared with a write voltage of around 10 V. A number of groups have therefore investigated the possibility of tunnelling through the oxide separating the poly1 poly2 layers in a standard, double polysilicon CMOS process. The structure developed by Thomsen and Brooke[89] is shown in figure A4. It consists of a sense MOSFET, a 100 fF coupling capacitor and a tunnelling injector.

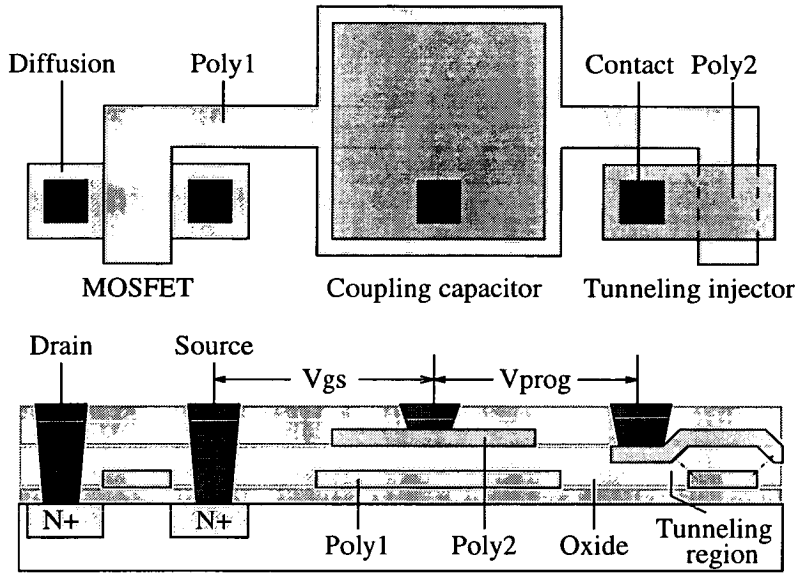
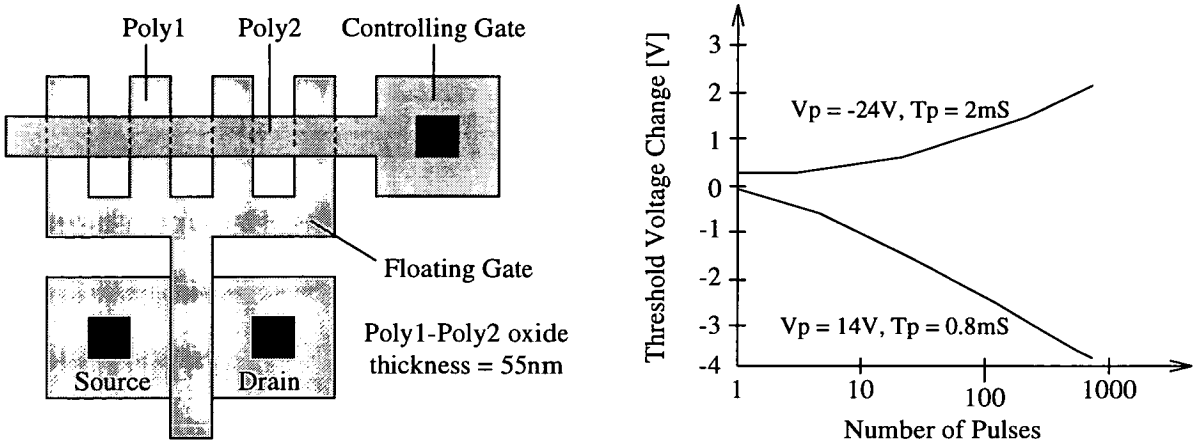


Figure A4 - Thomsen and Brookes standard CMOS floating gate cell

The injector is a poly1-poly2 structure; however, unlike a normal capacitor the upper polysilicon (poly2) overlaps the edge of the lower layer (poly1). The sharp edge of the poly1 slab causes field enhancement and thinning of the inter-polysilicon oxide. When a voltage of more than 12 V is applied to the structure bidirectional conduction occurs.

Again this device, based on a thicker oxide, has better charge retention characteristics than a conventional EEPROM. However, Thomsen and Brooke reported that it showed signs of aging after 1000 cycles, due to an accumulation of trapped electrons in the oxide. For this reason they suggested it might be more suited to trimming applications, rather than ones in which it will continually be written to.

Sheu[99] has also developed a poly1 poly2 tunnelling structure, this time for synaptic weight storage. The structure, shown in figure A5(a), uses the bump-like areas caused by poly2 overlap of poly1 to enhance the tunnelling field: the more bumps, the lower the tunnelling voltage.



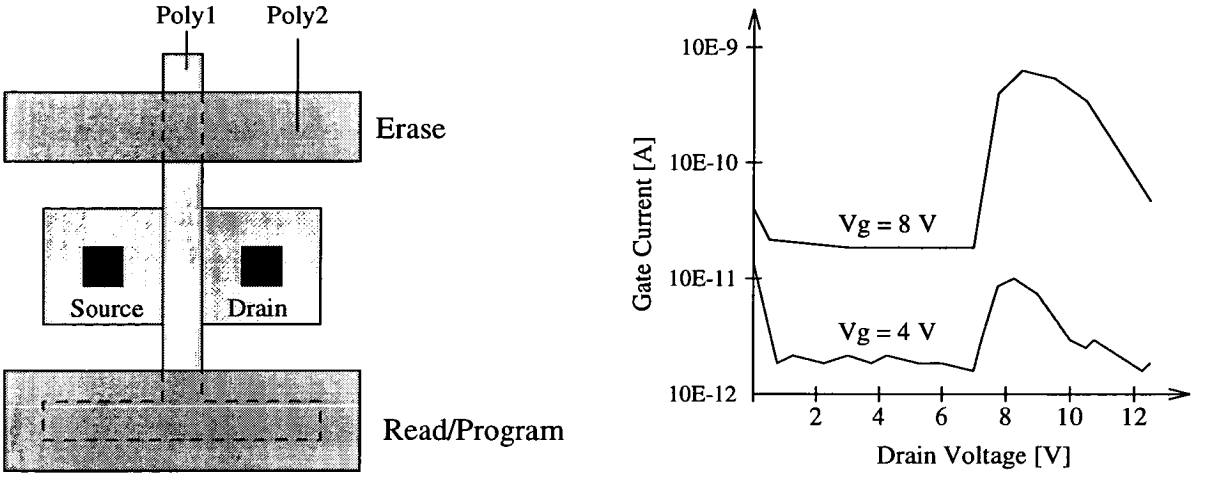
(a) Device layout (b) Programming results

Figure A5- Sheu's standard CMOS floating gate cell

The memory structure, based on the MOSIS 2 μm process, occupies an area of 60 μm x 70 μm. The disadvantage of this design is that a large number of high (24 V) millisecond programming pulses are required, as the programming characteristic in figure A5(b) illustrates.

In the literature on standard CMOS memory cells there is disagreement as to the role played by field enhancement and oxide thinning. Durfee and Shoucair[100] therefore designed a test chip containing a number of different injection capacitor structures. They found that the injector with poly2-poly1 overlap, and with the largest number of corners, could be programmed using voltage pulses as low as +6.5 V and -9 V, pulses typically 1 to 2 s in duration[101].

The final standard CMOS floating gate cell to be considered was designed by Montalvo[102]. By using a combination of hot-electron programming and Fowler-Nordheim erase he managed to achieve programming times in 100s μs and 10s ms erase times. The cell again relies on poly2-poly1 overlap, as shown in figure A6(a).



(a) Device layout (b) Programming results

Figure A6 - Montalvo's standard CMOS floating gate structure.

The cell contains two separate coupling capacitors, one for write and the other for erase. This allows each coupling capacitor to be tailored for its own particular role. For example, the process being used had an oxide breakdown of 16 V, and the minimum programming voltage was 6 V. The program coupling capacitor was therefore chosen so that K_w was greater than 40%. The memory cell is very compact occupying an area of only 12 μm x 17 μm, in 2 μm technology.

Figure A6(b) illustrates the effect of increased drain voltage on the tunnelling current: once the voltage exceeds the normal operating level (5 V) the tunnelling current increases sharply.

Conclusions

The programming characteristics of EEPROM cells make them particularly suitable for analogue storage. A commercial chip has been fabricated which uses 128,000 analogue EEPROM cells for real-time speech storage. The devices are specified for thousands of read/write cycles and the chip is available as a 5 V only part.

There has recently been significant research into floating gate cells constructed using standard CMOS processes. Although slower than standard EEPROM technology they are potentially much cheaper and also have better long term storage characteristics.

Appendix B

a-Si:H Device Fabrication

Introduction

To construct a-Si:H memory devices on the surface of a CMOS wafer a special fabrication procedure was developed in collaboration with the Department of Applied Physics and Electronics in Dundee, who carried out the fabrication. The final process sequence involved five different mask stages which are detailed below.

Wafer from ES2

The wafer from ES2 is covered in a passivation layer except over two structures: the bondpads and the contacts for the a-Si:H memory devices. Figure B1 below illustrates the passivation openings for the a-Si:H memory.

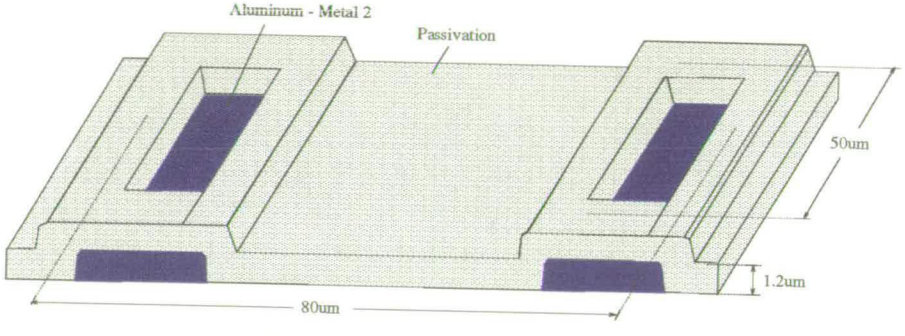


Figure B1 - Wafer from ES2

The height of the passivation shown in Figure B1 is 1.2 μm. This is the standard thickness of passivation for the ES2 1.5 μm process used for ASiTEST1. For the two subsequent designs a special arrangement was agreed with the ES2 Foundry for thin passivation, only 0.1 μm thick, an option that is only available if purchasing whole wafers.

Mask 1: Chromium Deposition

The bottom electrode of the a-Si:H memory device is chromium. This is deposited by d.c. sputtering to a thickness of 200Å.

The chromium is also used to cover exposed aluminium, namely the vanadium contact and the bondpads. This protective layer is needed as aluminium is used as the transfer layer when defining the a-Si:H layer.

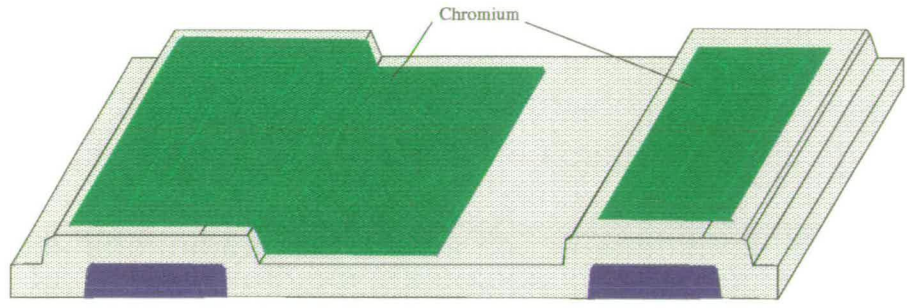


Figure B2 - Wafer after chromium deposition

Mask 2: a-Si:H Deposition

The deposition of the p-type a-Si:H:B is carried out by a Plasma Enhanced Chemical Vapour Deposition (PECVD) process. Silane gas (SiH_4) and the diborane (B_2H_6) dopant gas are introduced by a measured flow into the parallel plate PECVD reactor. One of these plates, the ground electrode, is heated to 220°C and holds the CMOS wafer substrate. The pressure inside the reactor is maintained at 0.1 torr by a pumping system. Using a few watts of radio frequency power, generally at 13.56 MHz, a weak plasma is produced between the reactor plates as the gas molecules break down. The a-Si:H:B alloy film then grows on the heated wafer substrate.

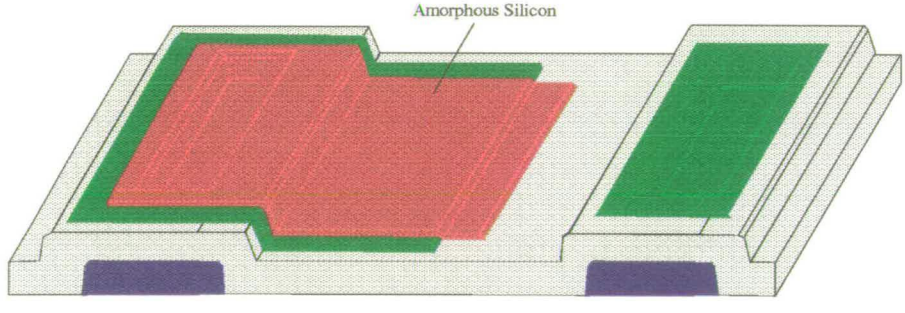


Figure B3 - Wafer after a-Si:H deposition

The advantages of this process are low temperature deposition, conformal coating, and that the CMOS circuits are not damaged in any way. Another advantage of a-Si is that it can be processed into devices using conventional photolithographic techniques.

Although the active area of the device is between the chromium/vanadium electrodes, the a-Si:H layer extends to the edge of the chromium electrode. This is intended to bridge any gaps caused by chromium fracturing around the steps in the passivation.

Mask 3: Active Pore Definition

The active pore of the device is defined using a layer of baked photoresist (Shipley S1818). The photoresist is deposited and selectively etched as normal. It is then baked at 200°C which hardens the photoresist so that it will not be etched away in subsequent

processes.

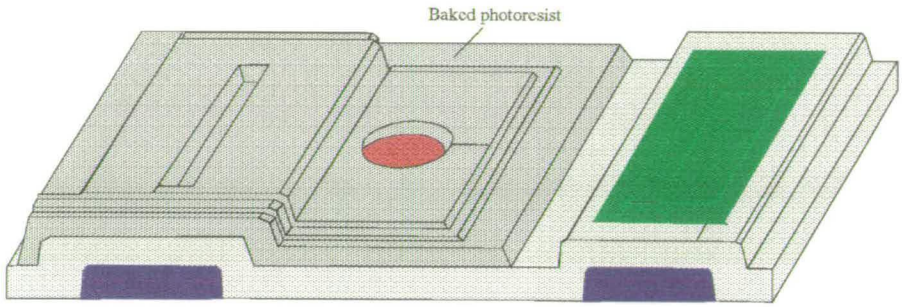


Figure B4 - Wafer after pore layer deposition

If this step is eliminated from the fabrication sequence then so called "overlap" memory devices, in which the active area is defined by the overlap of the vanadium on the a-Si:H with the bottom chromium contact, are created.

Mask 4: Vanadium Deposition

The top metal electrode is vanadium. This is deposited using d.c. sputtering to a thickness of 150Å.

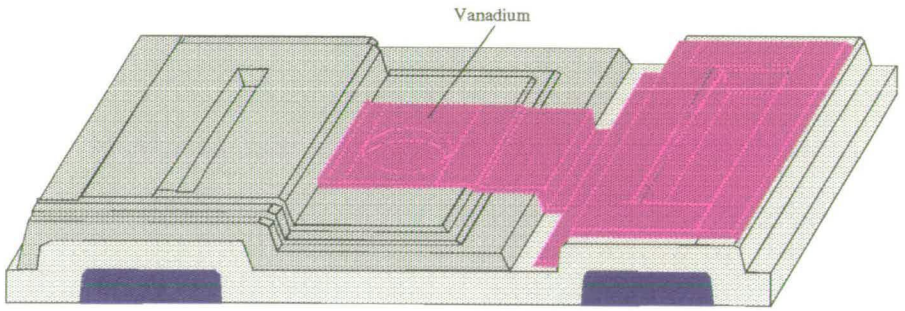


Figure B5 - Wafer after vanadium deposition

Mask 5: Cleaning the Bondpads

Although the a-Si:H memory device is now complete there is still a layer of protective chromium over the aluminium bondpads. The final mask is used to remove this chromium so that it is possible to wire bond to the aluminium of the bondpad.

Appendix C

Bonding and Pin Diagrams

Introduction

During this project three different chips were designed and fabricated. The chips were supplied by the manufacturer as half wafers. After the a-Si:H processing was complete the wafer segments were diced and a small number of parts wire bonded into Dual-In-Line (DIL) packages.

This appendix contains the different bonding and pin diagrams that were used during the project.

ASiTEST1

For the ASiTEST1 chip two different bonding configurations were required:

- Two terminal test structures in a 24-pin package
- FWE test blocks in a 40-pin package

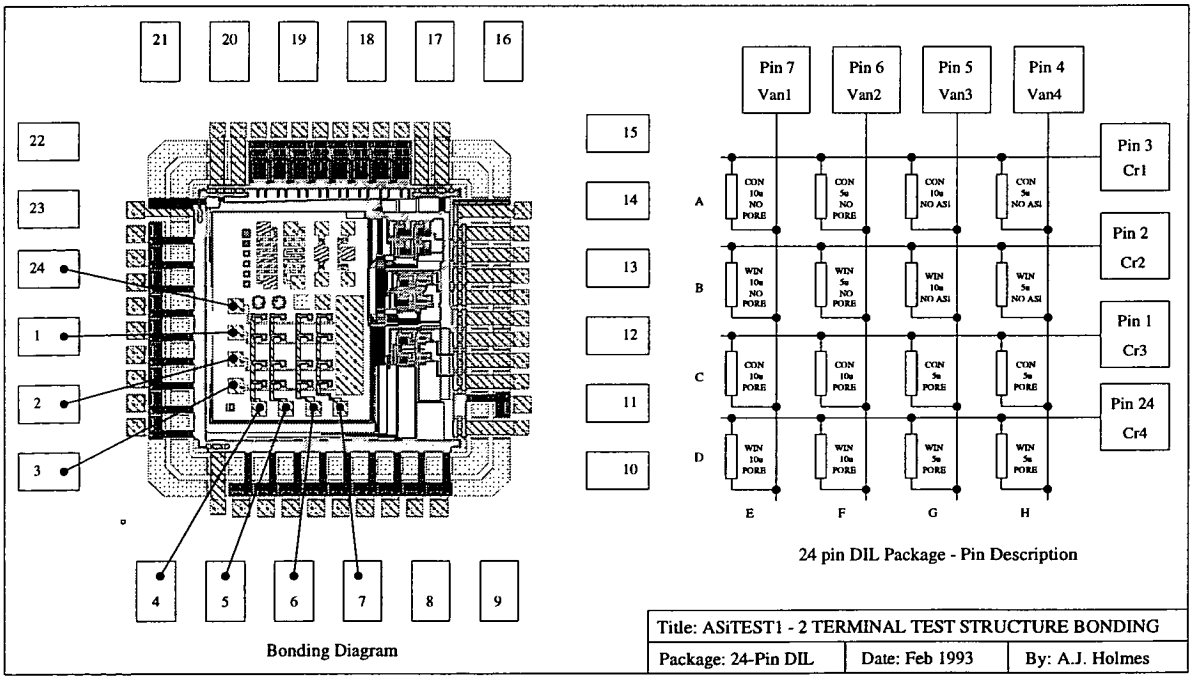


Figure C1 - ASiTEST1: Bonding for 2 Terminal Test Block

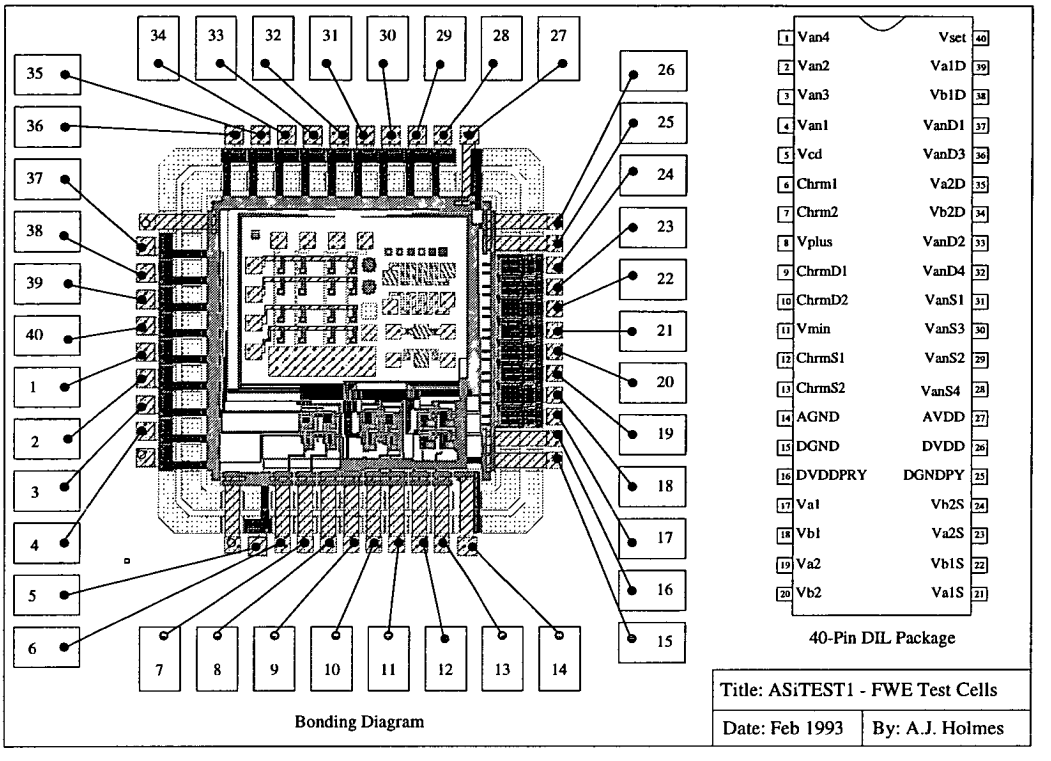


Figure C2 - ASiTEST1: Bonding for FWE Test Blocks

ASiTEST2

For the ASiTEST2 chip six different bonding configurations were required.

- Two 24-pin bonding configurations for the two-terminal test blocks.
- Four 40-pin configurations to test the different synapse designs.

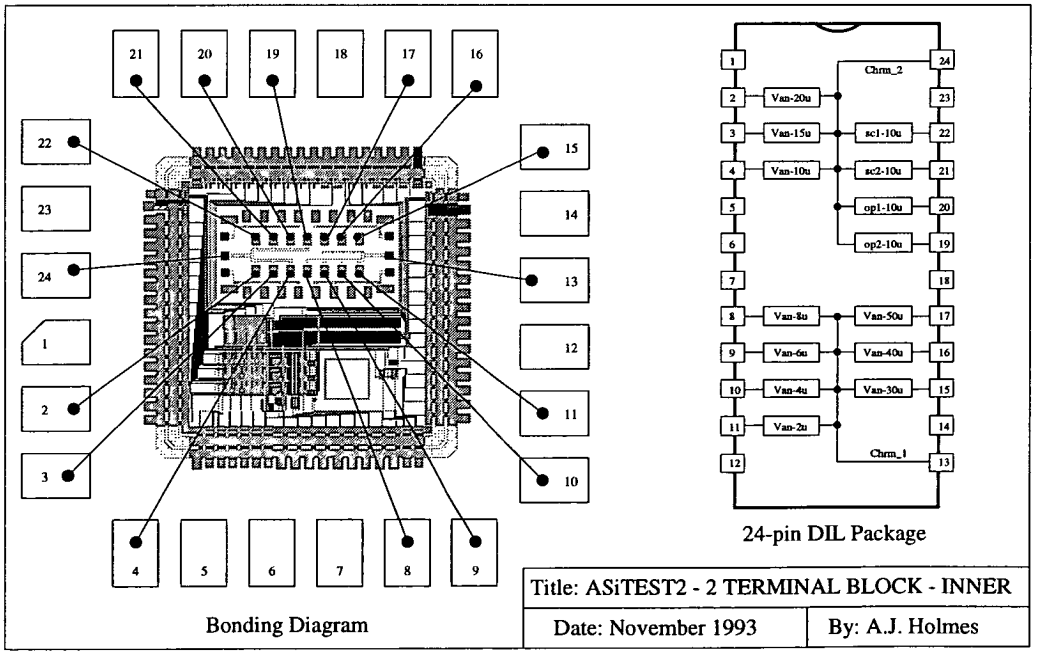


Figure C3 - ASiTEST2: Bonding for Inner set of Two Terminal Structures

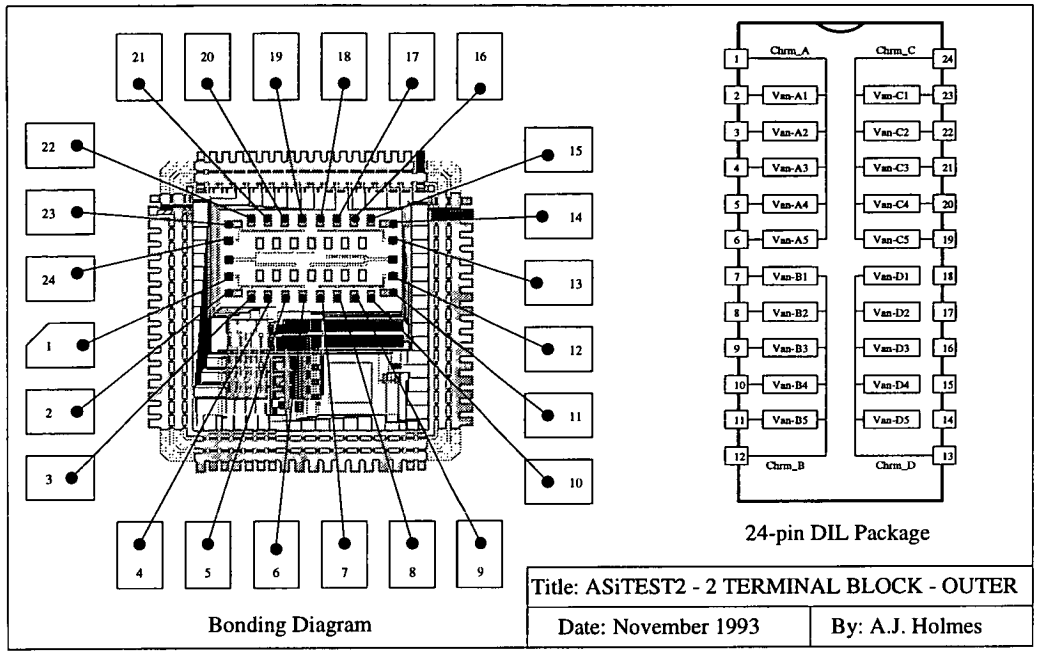


Figure C4 - ASiTEST2: Bonding for Outer set of Two Terminal Structures

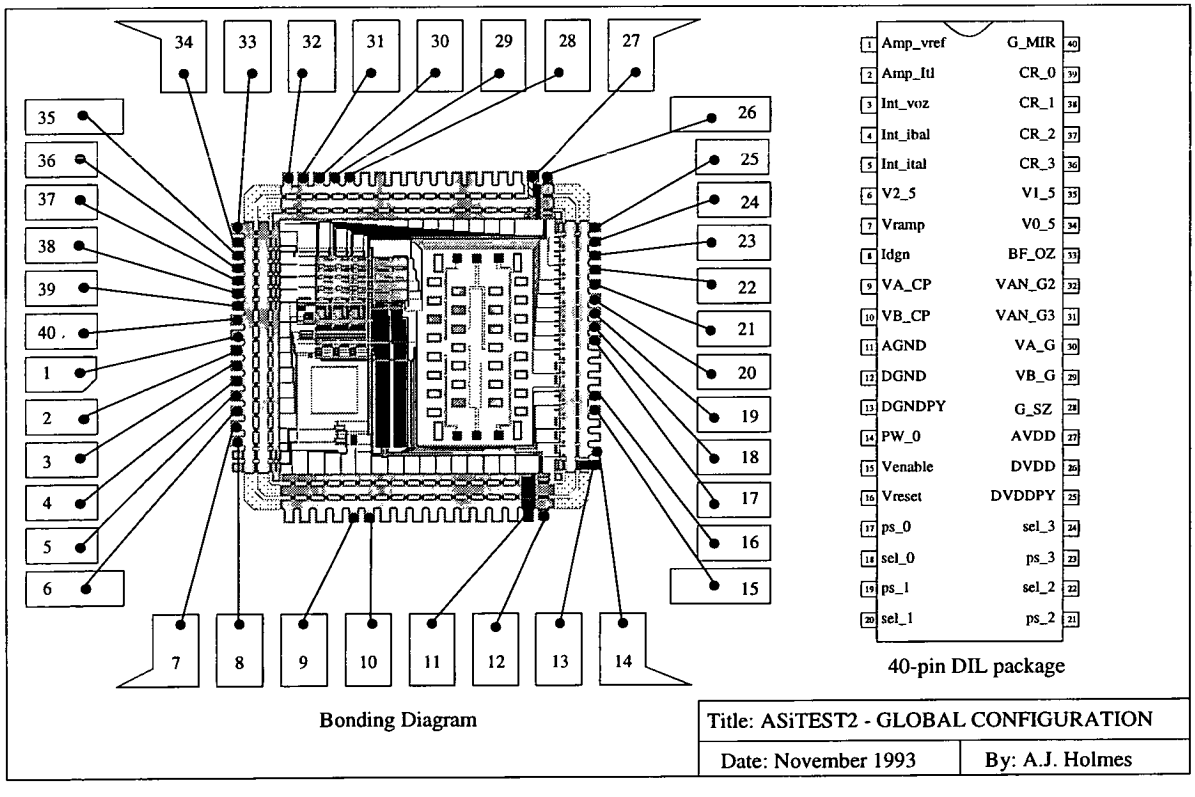


Figure C5 - ASiTEST2: Bonding for Global Synapse Test Block

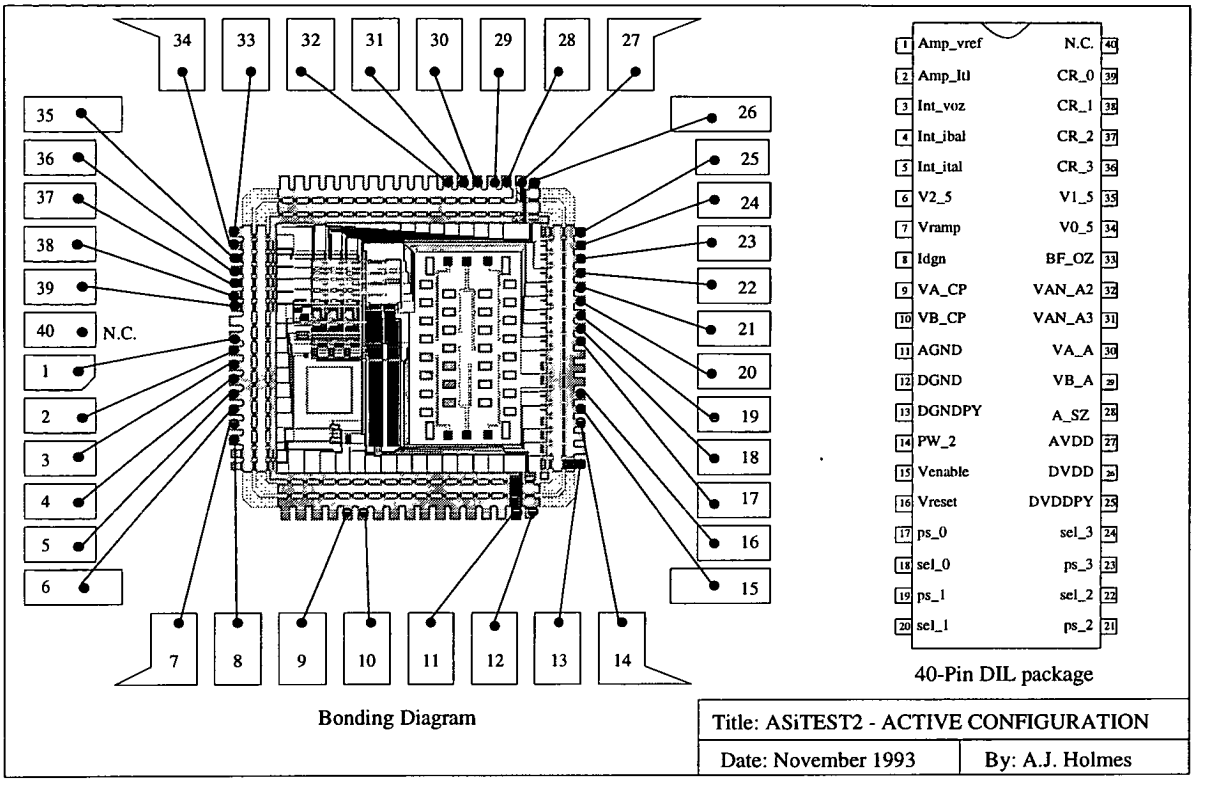


Figure C6 - ASiTEST2: Bonding for Active Synapse Test Block

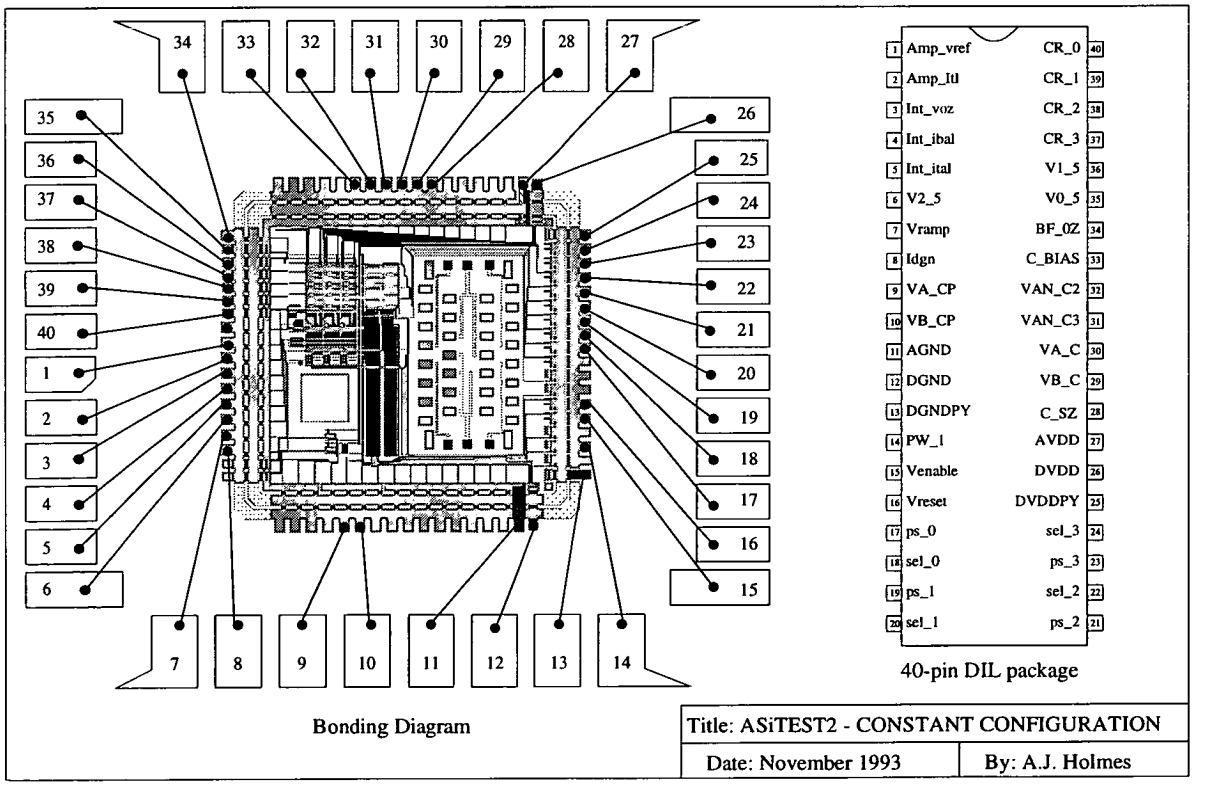


Figure C7 - ASiTEST2: Bonding for Constant Synapse Test Block

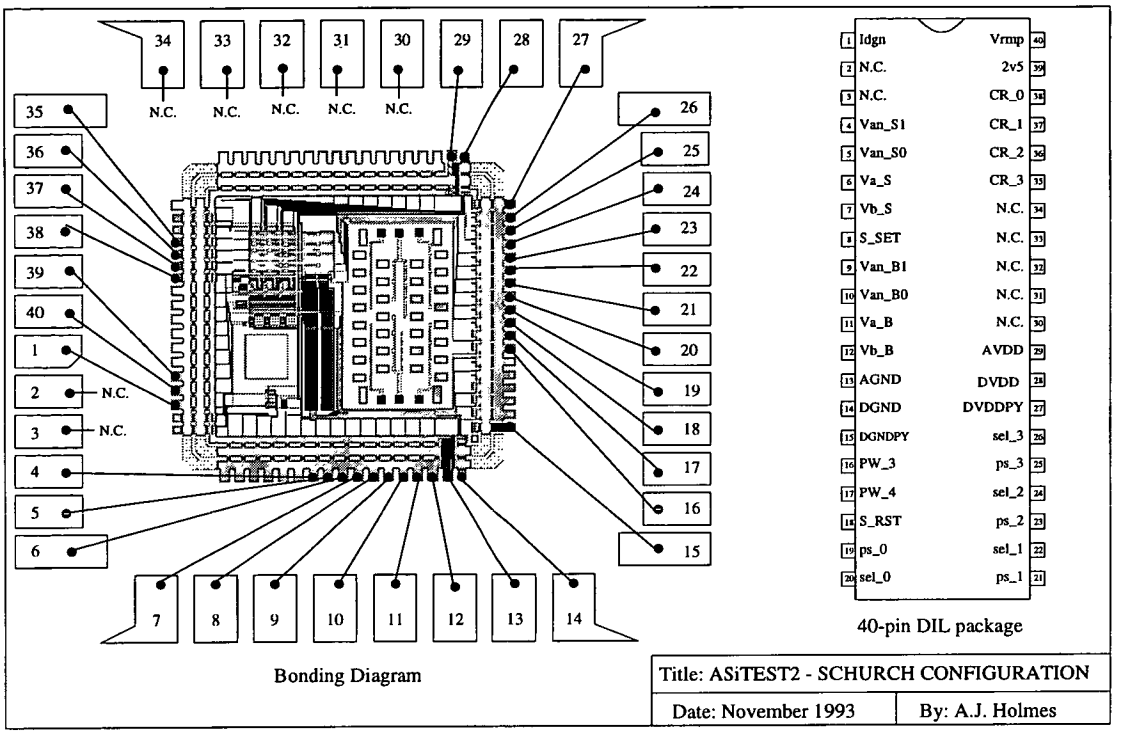


Figure C8 - ASiTEST2: Bonding for Schurch Synapse Test Block

ASiTEST 3

The ASiTEST3 chip only has one bonding configuration

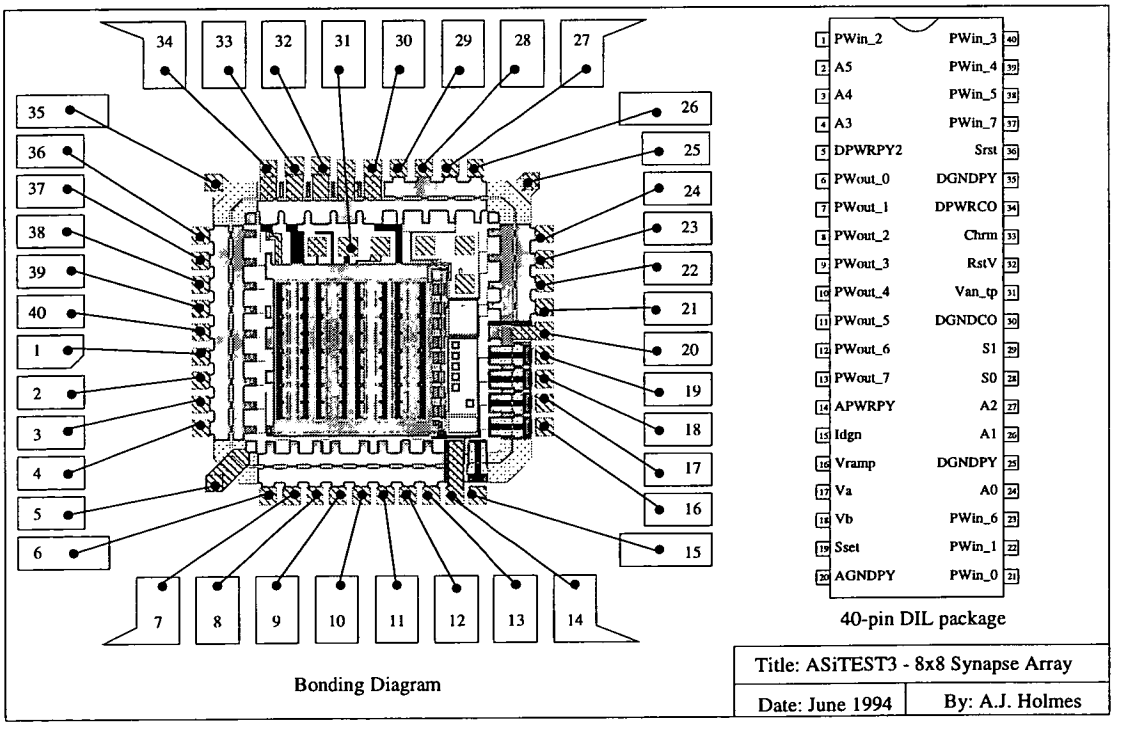


Figure C9 - ASiTEST3: Bonding for ANN with 8x8 Synapse Array

Appendix D

Test Equipment and Programmer Boards

Introduction

During the course of this project three different test chips were designed and fabricated, each with some form of CMOS circuitry accompanying a-Si:H memory devices. The chips also contained two terminal test structures with no CMOS circuitry. To test the CMOS portion of the chips a number of circuit boards were designed and constructed. The two terminal test structures were tested using the programmer circuitry already in the laboratory. This appendix briefly catalogues the different test boards used during the course of this project.

The original BBC Micro controlled set-up

The basic equipment needed to program and test a-Si:H memory devices includes a pulse generator, to apply write and erase pulses, and a means of measuring the device resistance after switching. The test setup in place at the start of this project is shown in figure D1.

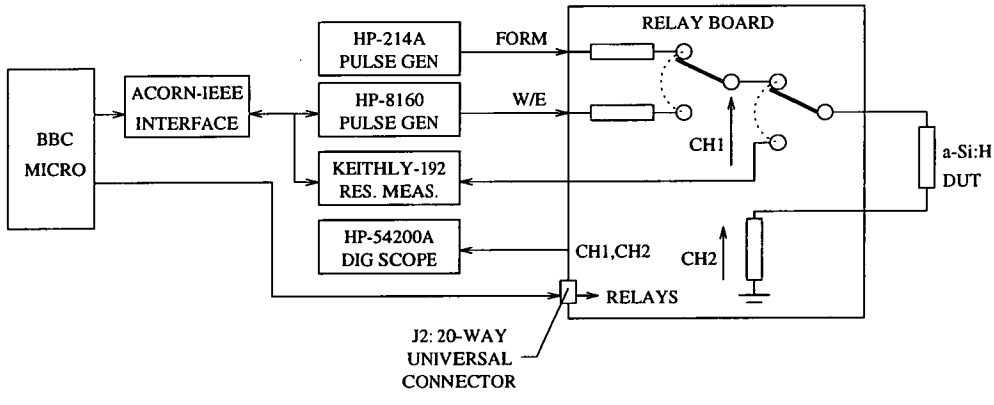


Figure D1 - Original programmer setup controlled by a BBC microcomputer

The BBC microcomputer controlled the various items of equipment using an Acorn-IEEE interface module.

Two terminal test board

The board used to perform all the two terminal programming experiments was one designed by a fourth year project student; it replaced a relay box originally used for programming. The circuit designed by the student, shown in figure D2, uses the same control

signals, generated by the BBC micro, as the original relay board.

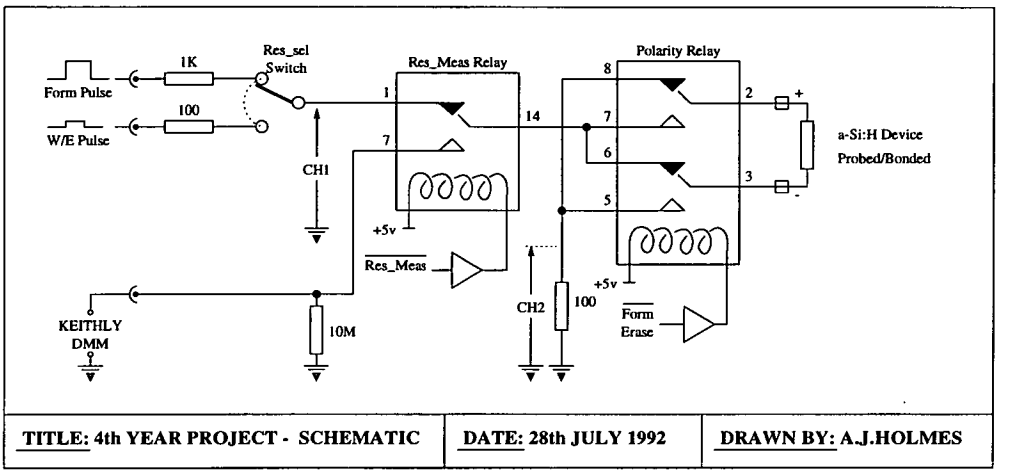


Figure D2 - Two-terminal programmer schematic

The programmer circuit was constructed on a small PCB, designed such that it fitted onto the connector panel of the Hewlett Packard oscilloscope, minimising the amount of cabling and hence reducing the amount of ringing on the fast programming pulses. The completed board is shown in figure D3.

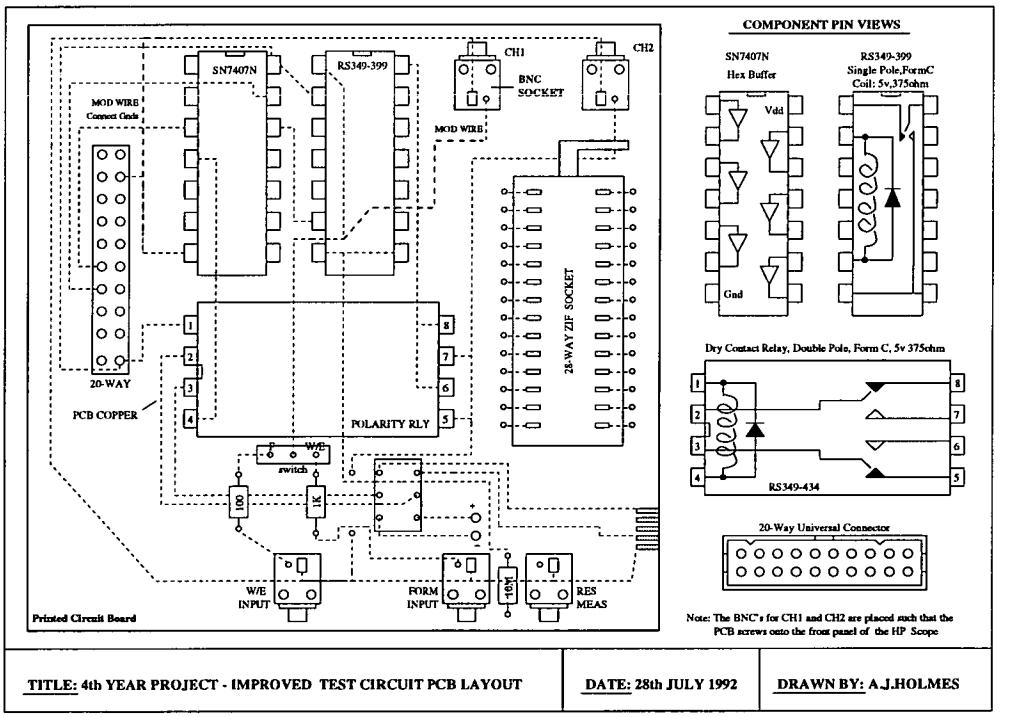


Figure D3 - Layout of two-terminal programmer board

PC-AT Controlled test setup - Hardware

The test system was originally controlled by a BBC micro computer. It was decided that this should be replaced with a a PC-AT fitted with a GPIB interface card. This offered the following advantages over the BBC:

i) The GPIB card could be used to control all the devices previously controlled by the BBC.

i) The PC could also be used to drive a prototype card containing various digital and analogue control circuits.

ii) The control software could be written in MicroSoft-C rather than BBC BASIC.

iii) Data files could be stored on 3.5 in diskettes rather than the aging 5.25 in disks used by the BBC.

A block diagram of the PC-AT controlled test setup is shown in figure D4.

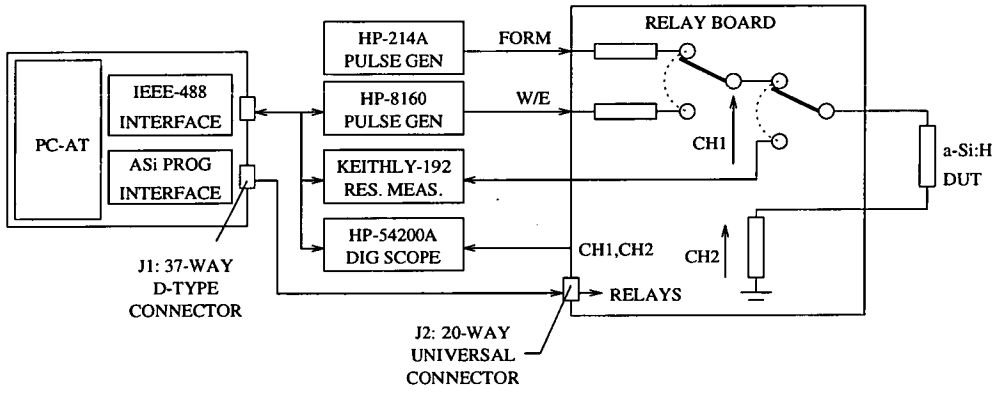


Figure D4 - Block diagram of test setup controlled by a PC-AT

PC-AT Controlled test setup - Software.

Software - Overview

Having replaced the BBC micro, software had to be written to control all the equipment linked by the GPIB network. This software was written in Microsoft C and was subdivided into a number of modules, one for each piece of hardware. Figure D4 shows a tree of the different software modules that were written.

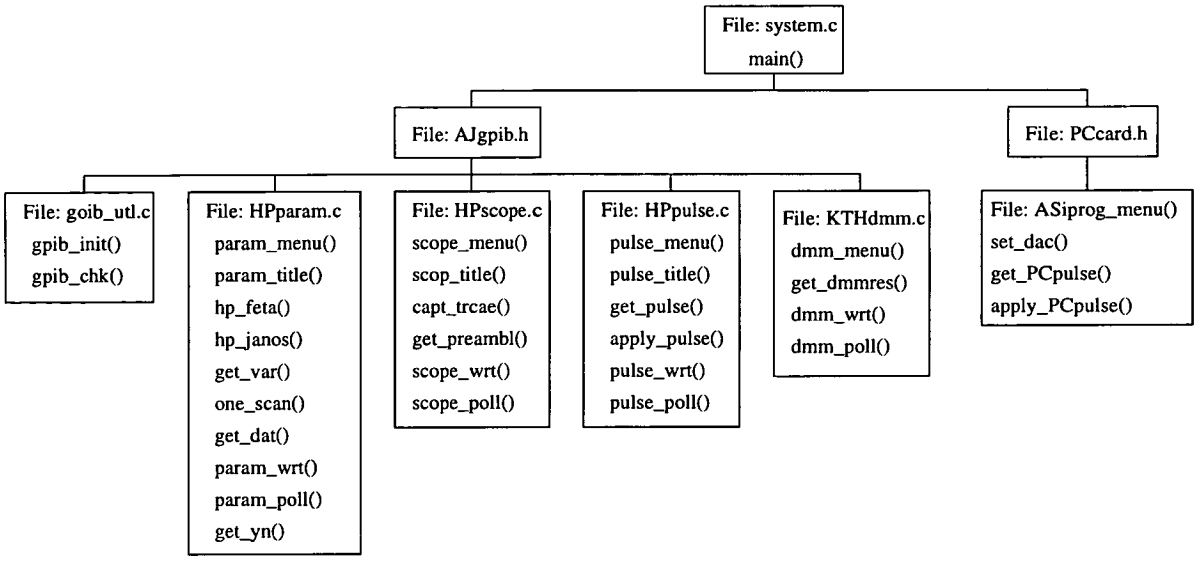


Figure D4 - Block diagram of the control software

This modular approach allowed new devices, such as the HP scope, to be easily incorporated into the existing GPIB software.

A separate library of C functions was also written to control the HP parameter analyser and Schlumberger frequency analyser. This allowed the results of various d.c. and a.c. analysis to be captured and stored on diskette for the first time.

The various control programs all use a common display format. A text window is used to display menus and runtime messages. A graphics window is used to display results and scope traces.

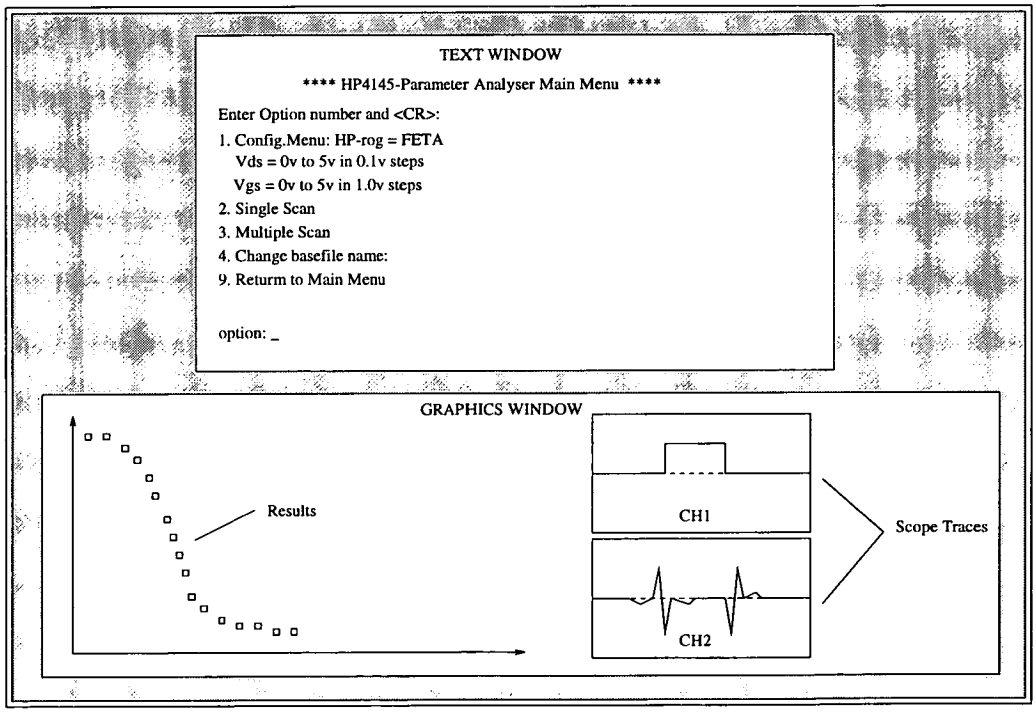


Figure D5 - Control software graphics display

ASiTEST test boards

Three different chips were designed during of this project and each required a test board to supply addressing and programming signals. Each of the three chips will now be considered briefly in turn.

ASiTEST1

On the ASiTEST1 chip there were three different FWE cells. As they all required similar addressing signals it was decided that a single test board containing one set of addresser circuitry should be built. To test a particular FWE cell a DIL header could then be used to connect the appropriate pins on the chip to the control signals. Figure D6 is a schematic of the ASiTEST1 board and figure D7 is the layout of the test board.

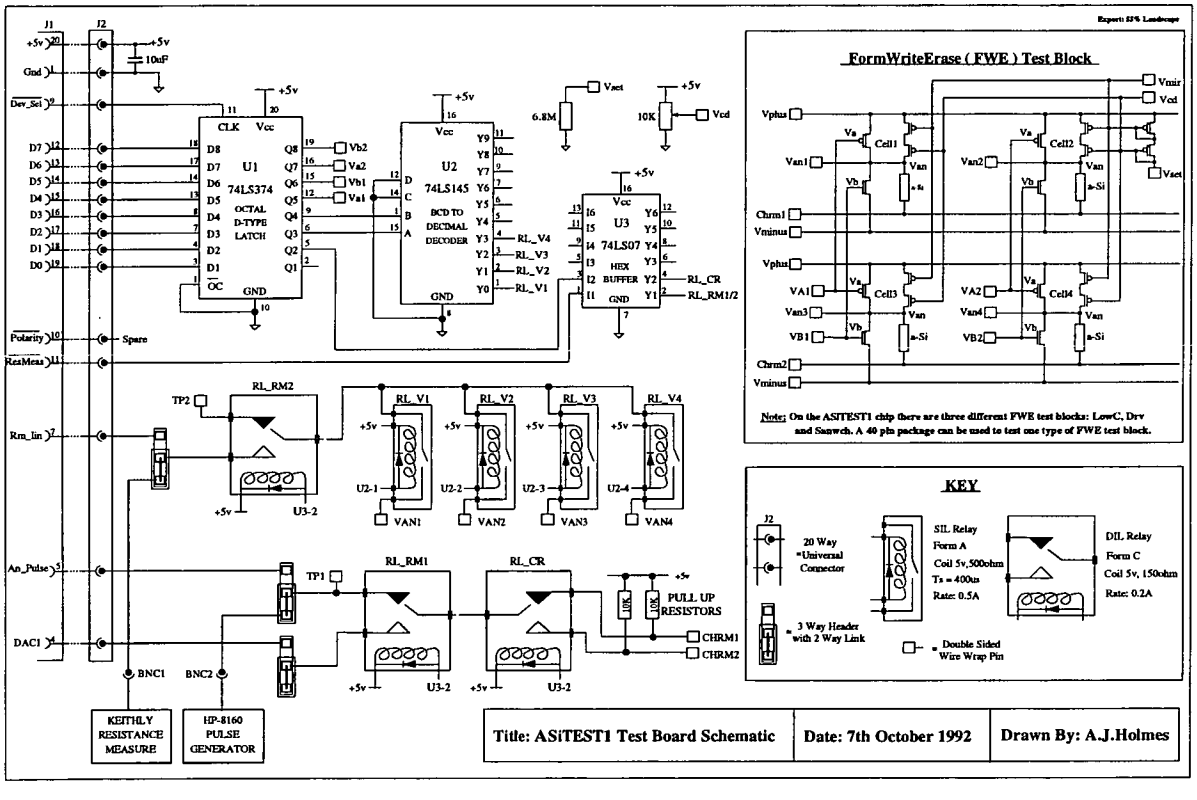


Figure D6 - ASiTEST1 Board Schematic

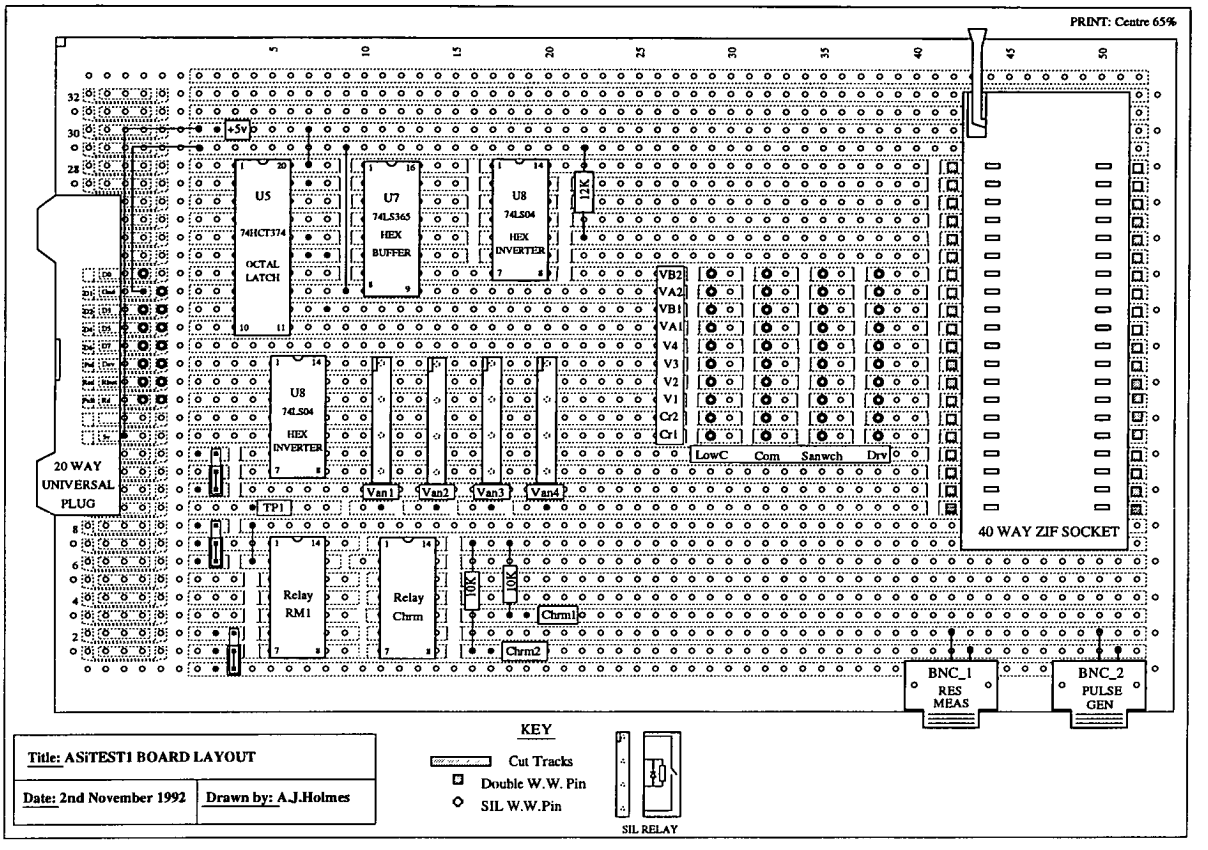


Figure D7 - ASiTEST1 Board Layout

ASiTEST2

On the ASiTEST2 chip there were five different synapse designs. Three of them were based on the EPSILON synapse and two on the Schurch one. It was therefore decide that two test boards should be built. The first tested the three EPSILON designs using a header arrangement similar to that on the ASiTEST1 board. The second board was designed to test both Schurch designs at the same time.

To test the synapse performance the output pulsewidth had to be recorded. This was done using a DAC with a built in counter. When the ramp signal reached the same level as the integration capacitor on the chip the PWout signal goes high and the DAC counter is stopped. The counter value can then be read back into the PC.

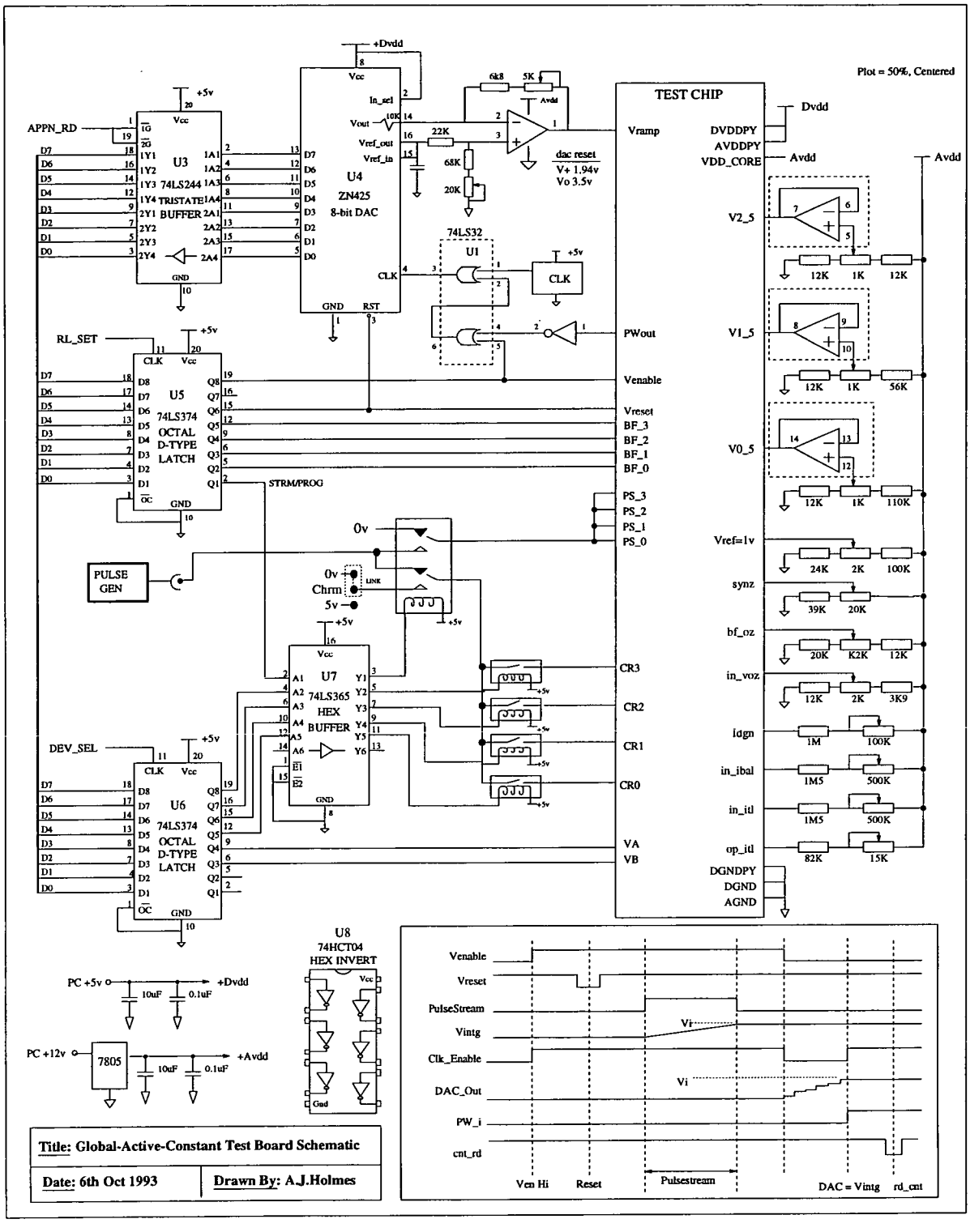


Figure D8 - ASiTEST2 Board 1 schematic

As the second test board is designed to test two blocks of synapses at the same time it contains an additional select line to determine which of the PWout signals is the one that stops the DAC counter.

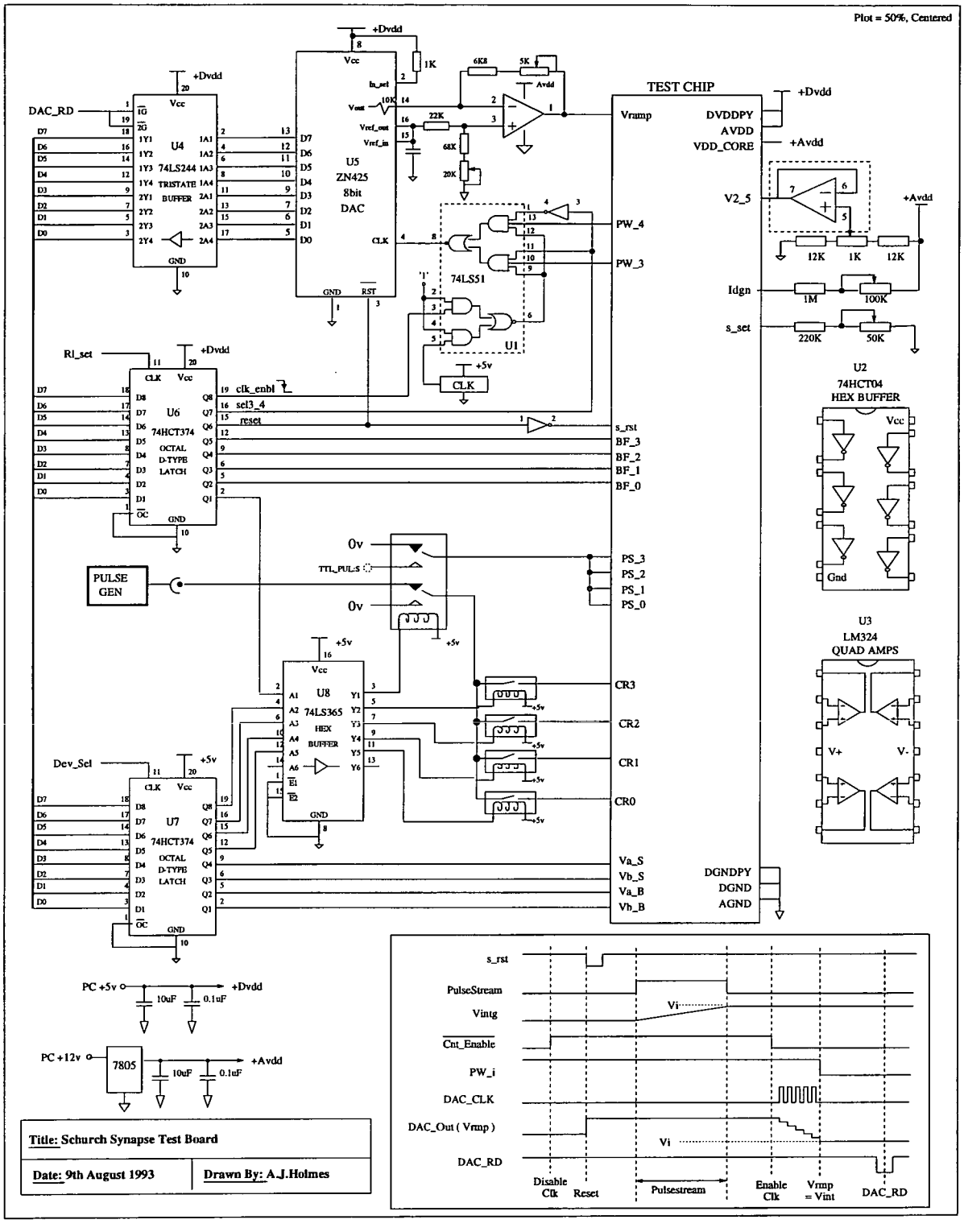


Figure D9 - ASiTEST2 Board 2 schematic

Figures D10 and D11 contain the layout diagrams for the ASiTEST2 boards.

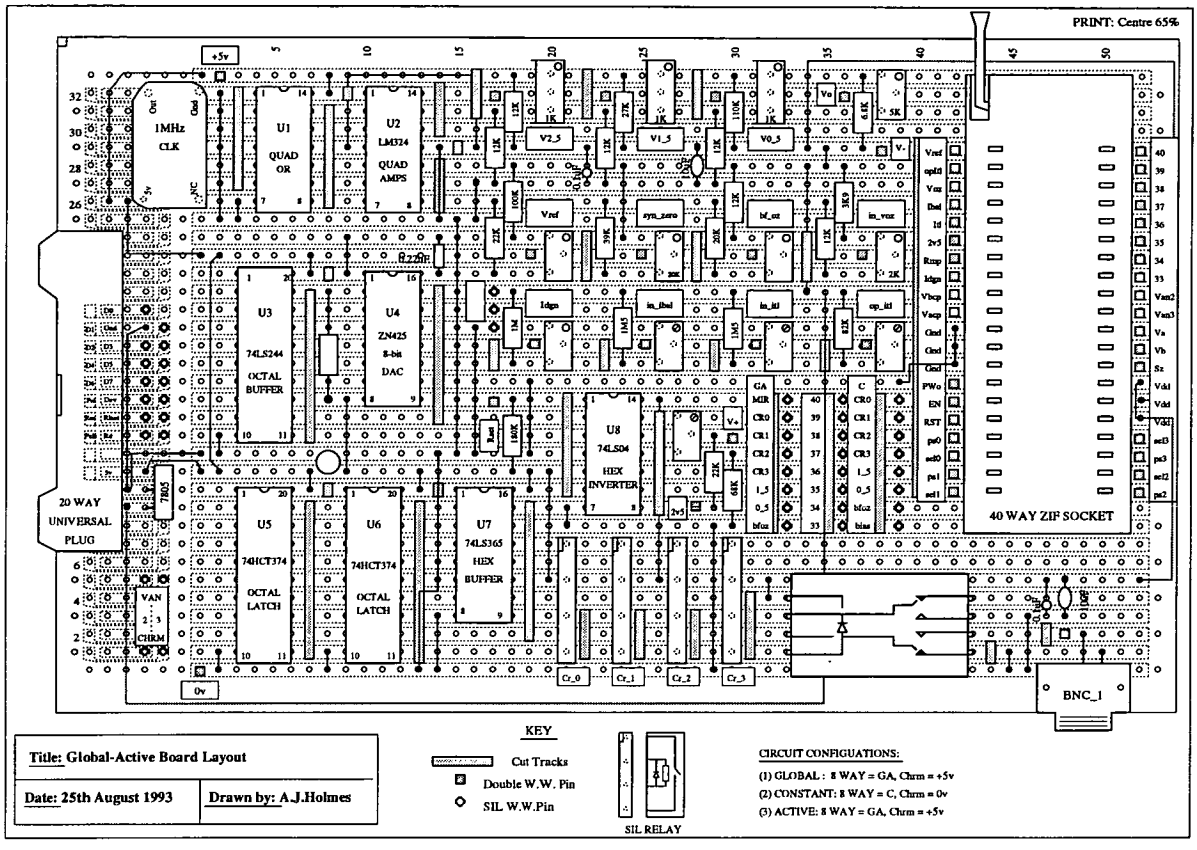


Figure D10 - ASiTEST2 Board 1 layout

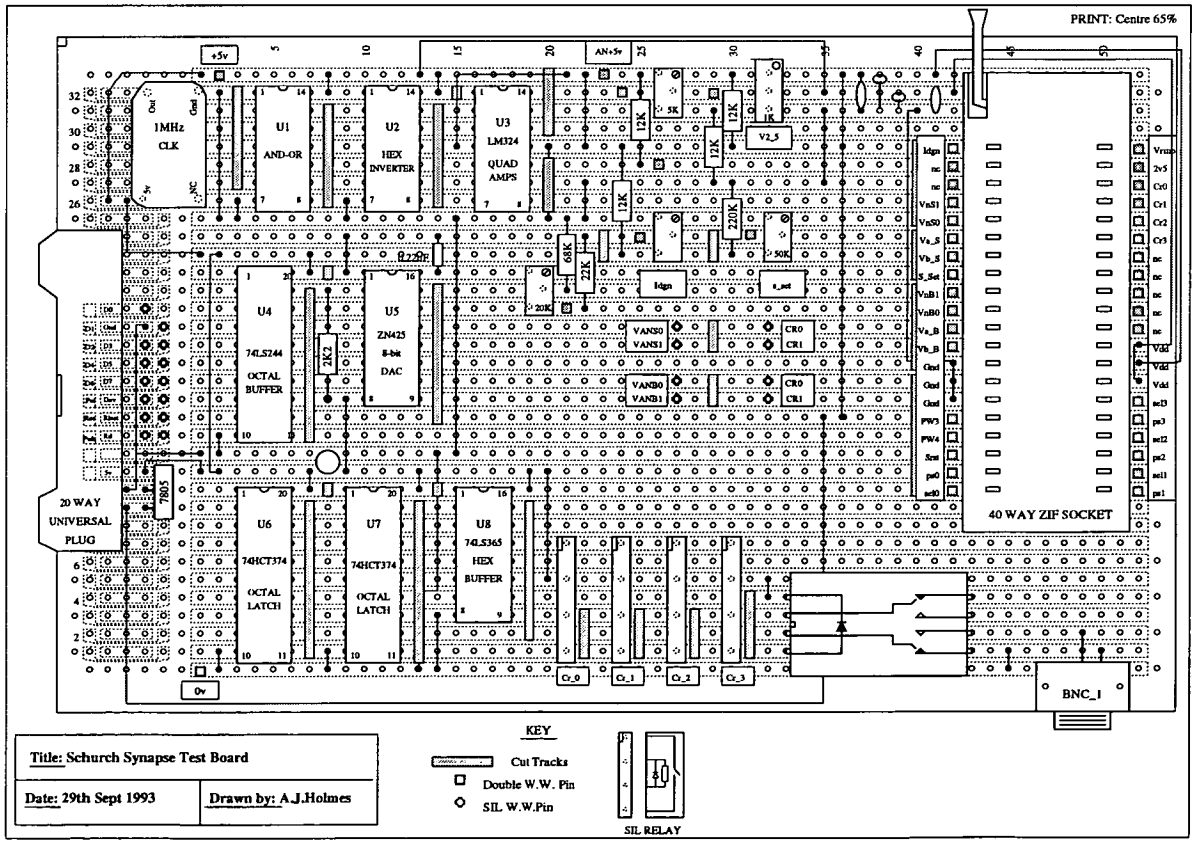


Figure D11 - ASiTEST2 Board 2 layout

ASiTEST3

The design of the ASiTEST3 board is discussed in chapter 5. It contains a simple state machine and is designed to run two ASiTEST3 chips cascaded in series.

Generation of PWin signal

On the original EPSILON test board the PWin signals were generated by loading the required bit pattern into 256 locations in SRAM. The output signal was then generated by clocking through the SRAM, as illustrated in figure D12.

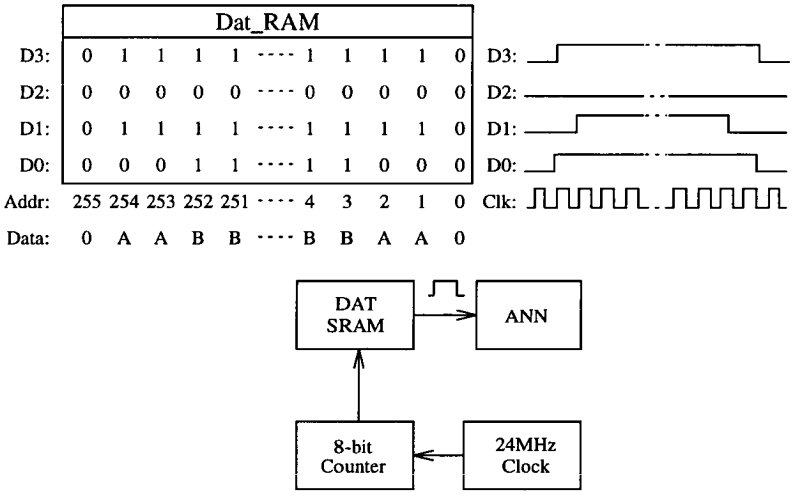


Figure D12 - Original PWin generation Scheme

On the EPSILON test board this need to download 256 bytes for every input pattern has considerably slowed the cycle time needed for a forward pass. It was therefore decided that a different scheme should be used on ASiTEST3. Rather than downloading the value of the pulsewidth inputs for every clock count, 0 to 256, the new scheme only requires that they are loaded at transition times. This means that for a chip with eight inputs we only need to download 8 bytes. An extra RAM chip is used to store the transistion times. When the value on the eight bit counter is equal to the value stored in this Count_Ram the address counter is clocked, so placing a new set of PWin signals onto the Dat_Ram outputs.

The counter scheme has also been changed so that it counts 0 to 128 0, instead of 0 to 256. This is because every pulsewidth signal is assumed to be symmetrical.

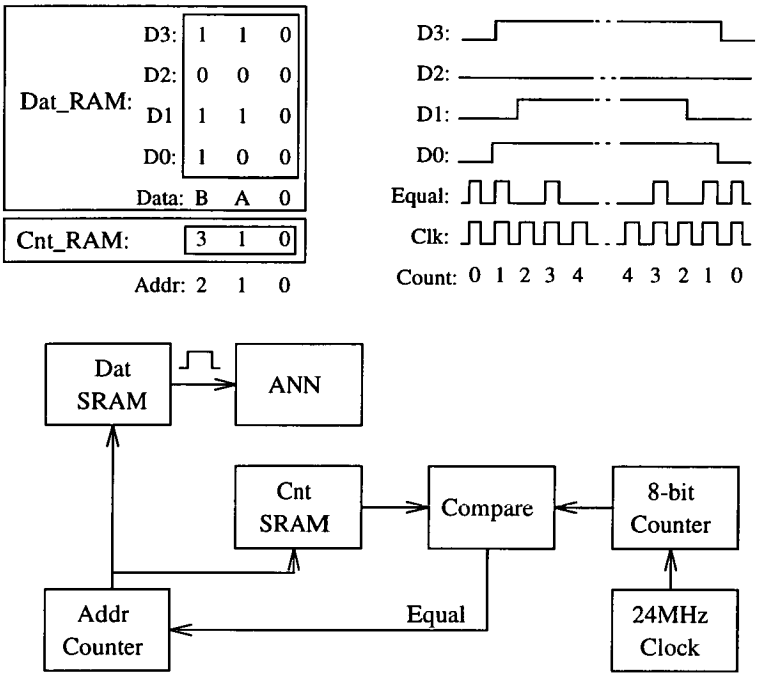


Figure D13 - ASiTEST3 PWin generation Scheme

The layout of the ASiTEST3 board is shown in figure D14 and the schematic in figure D15.

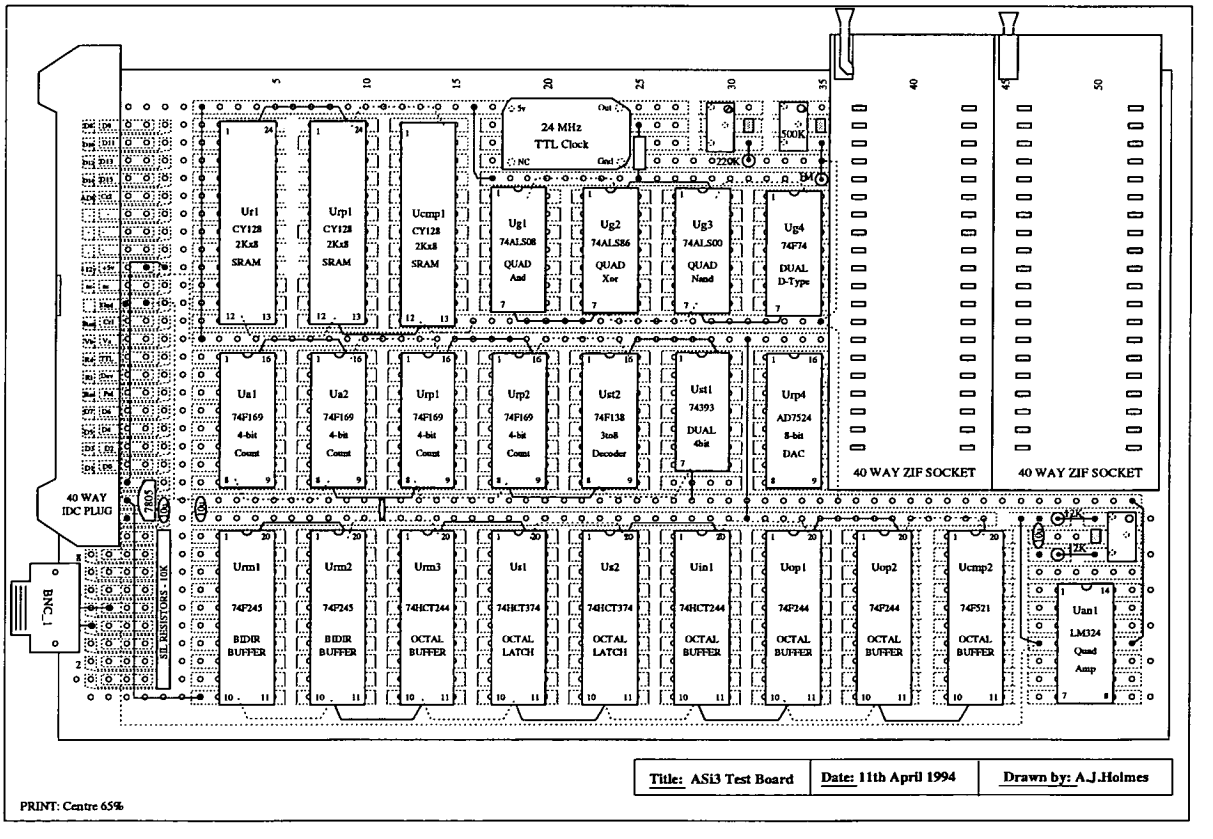


Figure D14 - ASiTEST3 Board layout

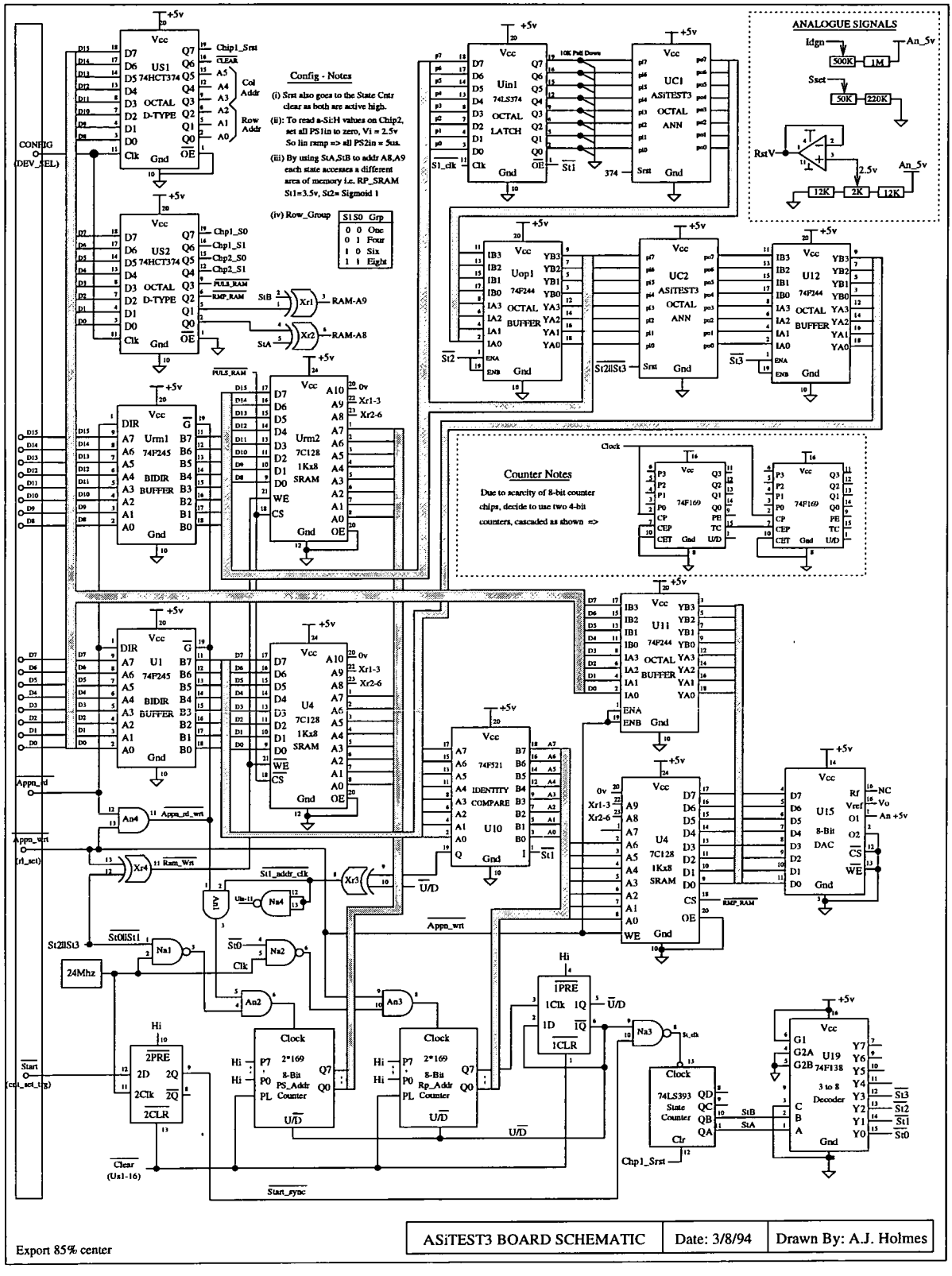


Figure D15 - ASiTEST3 Board schematic

Appendix E

HSPICE Modelling of a-Si:H switching

HSPICE Modelling of a-Si:H switching

In section 3.3.3 the I-V characteristic used to simulate the switching behaviour of the a-Si:H memory device was discussed. This was implemented in HSPICE as a Current Controlled Voltage Source (CCVS), with the I-V relationship being defined by a set of current/voltage pairs taken from actual device measurements. This model was then included in an HSPICE library element, as illustrated below.

```
.LIB ASi1_760
CASi1 van_1 cr_1 1e-12          $ a-Si:H capacitance
Vcr_1 van_1 n1 0v               $ Zero potential voltage source
HaSi1 n1 cr_1 PWL(1) Vcr_1      $ CCVS model of aSi:H device
+   -4.500e-3 , -3.900e+0   -3.500e-3 , -3.063e+0   -3.100e-3 , -2.689e+0
+   -2.950e-3 , -1.859e+0   -2.900e-3 , -1.820e+0   -1.000e-4 , -6.800e-2
+   -5.000e-5 , -3.500e-2    0.000e+0 ,  0.000e+0    5.000e-5 ,  3.500e-2
+    1.000e-4 ,  6.800e-2    2.900e-3 ,  1.820e+0    2.950e-3 ,  1.859e+0
+    3.000e-3 ,  1.971e+0    3.050e-3 ,  2.524e+0    3.100e-3 ,  2.689e+0
+    3.450e-3 ,  3.021e+0    3.500e-3 ,  3.063e+0    4.500e-3 ,  3.900e+0
.ENDL ASi1_760
```

During the design of the CMOS addresser circuitry two a-Si:H models were used, one low resistance (760 Ω) and one high resistance (604 kΩ). The I-V pairs used to construct these models are shown in Figure E1.

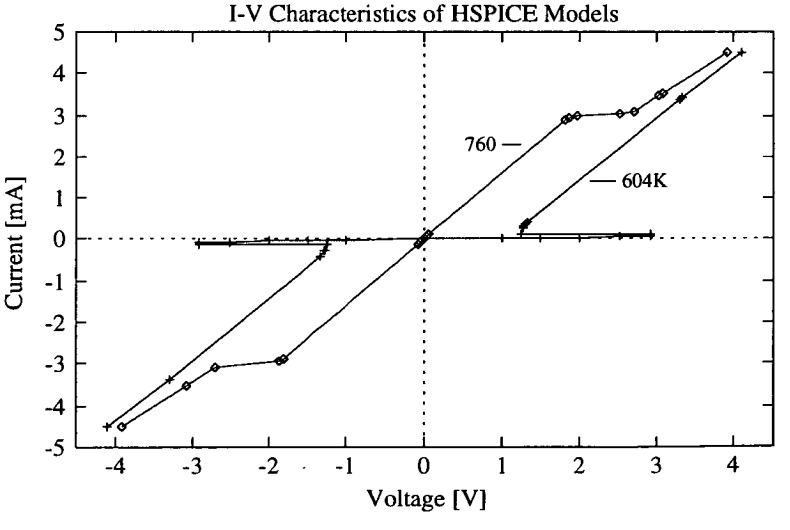


Figure E1 - Data sets used to construct 760 and 604K Models

Appendix F

References

References

1. R.P. Lippmann, "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, pp. 4 - 22, April, 1987.
2. E. Sackinger, B.E. Boser, J. Bromley, Y. Lecun, and L.D. Jackel, "Application of the ANNA Neural Network Chip to High-Speed Character Recognition", *IEEE Transactions on Neural Networks*, vol. 3, no. 3, pp. 498-505, 1992.
3. P.H.W. Leong and M.A. Jabri, "Kakadu - A Low-Power Analog Neural Network Classifier", *International Journal of Neural Systems*, vol. 4, no. 4, 1993.
4. A.F. Murray, A. Hamilton, D.J. Baxter, S. Churcher, H.M. Reekie, and L. Tarassenko, "Integrated Pulse-Stream Neural Networks - Results, Issues and Pointers", *IEEE Trans. Neural Networks*, pp. 385-393, 1992.
5. Y. Tsividis and S. Satyanarayana, "Analog Circuits for Variable Synapse Electronic Neural Networks", *Electronics Letters*, vol. 23, no. 24, pp. 1313-1314, 1987.
6. E. Vittoz, H. Oguey, M.A. Maher, O. Nys, E. Dijkstra, and M. Chevroulet, "Analog Storage of Adjustable Synaptic Weights", *Proc. ITG/IEEE Workshop on Microelectronics for Neural Networks, Dortmund (Germany).*, pp. 69-79, June 1990.
7. B. Hochet, "Multivalued MOS Memory for Variable Synapse Neural Networks", *Electronics Letters*, vol. 25, no. 10, pp. 669-670, 1989.
8. P. Hasler and L. Akers, "A Refreshable Multilevel Memory for a Continuous-Time Synapse", *Proc. 1992 IEEE International Symposium on Circuits and Systems (ISCAS -92)*, vol. 6, pp. 1561-1564.
9. Y. Horio, M. Ymamamoto, and S. Nakamura, *Active Analog Memories for Neuro-Computing*, 4, pp. 2986-2989.
10. B. Linares-Barranco, E. Sanchez-Sinencio, A. Rodriguez-Vasquez, and J.L. Huertas, "A CMOS Analog Adaptive BAM with On-Chip Learning and Weight Refreshing", *IEEE Transactions on Neural Networks*, vol. 4, no. 3, pp. 445-455, 1993.

11. D. Macq, J.D. Legat, and P.G.A. Jespers, "Analog Storage of Adjustable Synaptic Weights", *Applications of Neural Networks III - Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE)*, vol. 1709, pp. 712-718, 1992.

12. A.A. Reeder, I.P. Thomas, C. Smith, J. Wittgreffe, D.J. Godfrey, J. Hajto, A.E. Owen, A.J. Snell, A.F. Murray, M.J. Rose, and P.G. LeComber, "Application of Analogue a-Si Memory Devices to Resistive Synapses for Neural Networks", *Proc. International Conference on Microelectronics for Neural Networks (Munich)*, pp. 253-260, 1991.

13. M.J. Rose, A.J. Snell, P.G. LeComber, J. Hajto, A.G. Fitzgerald, and A.E. Owen, "Aspects of Nonvolatility in a-Si-H Memory Devices", *Amorphous Silicon Technology*, vol. 258, pp. 1075-1080, 1992.

14. C. Hu, *Non-Volatile Semiconductor Memories - Technologies, Design and Applications,* pp. 1-2, New York IEEE Press, 1991.

15. G. McLintock, "Nonvolatile Memory-Systems Move in New Directions", *Computer Design*, vol. 27, no. 22, p. 77, 1988.

16. J. Reimer, "Memories in My Pocket", *BYTE*, vol. 16, no. 2, pp. 251-258, 1991.

17. D. Kahng and S. Sze, "A Floating Gate and its Application to Memory Devices", *Bell Systems Technical Journal*, vol. 46, pp. 1288-1295, 1967.

18. D. Frohman-Bentchkowsky, "A Fully Decoded 2048-Bit Electrically Programmable FAMOS Read-Only Memory", *IEEE Journal of Solid State Circuits*, vol. 6, no. 5, pp. 301-306, 1971.

19. M. Woods, "An E-PROMs Integrity Starts with its Cell Structure", *Electronics*, August 1980.

20. W. Johnson, G. Perlegos, A. Renniger, G. Kuhn, and R. Ranaganath, "A 16KB Electrically Eraseable Non-Volatile Memory", *IEEE ISSCC Digest Technical Papers*, vol. 271, pp. 152-153, 1980.

21. Y. Terada, K. Kobayashi, T. Nakayama, M. Hayashikoshi, and Y. Miyawaki, "120ns 128k x 8b CMOS EEPROMS", *IEEE Journal of Solid State Circuits*, vol. 24, no. 5, pp. 1244-1249, 1989.

22. S.K. Lai, V.K. Dham, and D. Guterman, "Comparison and Trends in Today's Dominant EE Technologies", *IEDM Technical Digest*, pp. 580-583, 1986.

23. B.C Cole, "How SEEQ is Pushing EEPROMS to 1-MB Densities", *Electronics*, vol. 59, no. 29, pp. 53-56, 1986.

24. S. Weber, "Look Out EPROMS, Here Comes Flash", *Electronics*, vol. 63, no. 11, p. 44, 1990.

25. "Solid state recording for camcorder", *Electronics and Wireless World*, vol. 100, no. 1704, p. 886, 1994.

26. "Catalyst EEPROM Needs a Miserly 3 Volts", *Electronics*, vol. 60, no. 14, pp. 67-68, 1987.

27. B.C. Cole, "How the United States is Leading the Way in Strategic Nonvolatile Technology", *Electronics*, vol. 62, no. 3, pp. 30-33, 1989.

28. M. Bloom, "A Memory to Remember", *Electronics System Design Magazine*, October 1989.

29. D. Frohman-Bentchkowsky, "The Metal-Nitride-Oxide-Silicon (MNOS) Transistor - Characteristics and Applications", *Proceedings of the IEEE*, vol. 58, no. 8, pp. 1207-1219, 1970.

30. W.R. Iversen, "SIMTEK Revives SNOS for Nonvolatility", *Electronics*, vol. 62, no. 4, pp. 30-32, 1989.

31. M.H. White and C.Y. Chen, "Electrically Modifiable Nonvolatile Synapses for Neural Networks", *Proc. 1989 IEEE International Symp on Circuits and Systems (ISCAS-89)*, pp. 1213-1216.

32. T. Hagiwara, Y. Yatsuda, R. Kondo, S.I. Minami, and T. Aoto, "A 16 kBit Electrically Eraseable PROM Using N-Channel Si-Gate MNOS Technology", *IEEE Journal of Solid State Circuits*, vol. sc-15, no. 3, pp. 346-353, 1980.

33. J.L. Moll and Y. Tarui, "A New Solid State Memory Resistor", *IEEE Transactions on Solid State Devices*, vol. 10, pp. 338-339, 1963.

34. J.T. Evans and R. Womack, "An Experimental 512-Bit Non-volatile Memory with Ferroelectric Storage Cell", *IEEE Journal of Solid State Circuits*, vol. 23, pp. 1171-1175, 1988.

35. R. Moazzmi, C. Hu, and W.H. Shepherd, "A Ferroelectric DRAM Cell for High-Density NVRAM's", *IEEE Electron Device Letters*, vol. 11, no. 10, pp. 454-456, 1990.

36. D.W. Greve, "Programming Mechanism of Polysilicon Resistor Fuses", *IEEE Transactions on Electron Devices*, vol. ED-29, no. 4, pp. 719-724, 1982.

37. V. Malhotra, J.E. Mahan, and D.L. Ellsworth, "Fundamentals of Memory Switching in Vertical Polycrystalline Silicon Structures", *IEEE Transactions on Electron Devices*, vol. ED-23, no. 11, pp. 2441-2449, 1985.

38. J. Greene, E. Hamdy, and S. Beal, "Antifuse Field Programmable Gate Arrays", *Proceedings of the IEEE*, vol. 81, no. 7, pp. 1042-1056, 1993.

39. B. Cook and S. Keller, "Amorphous Silicon Antifuse Technology", *IEEE Bipolar Circuits Technol. Meet.*, pp. 99-100, October 1986.

40. H.P. Graf and L.D. Jackel, "Analog Electronic Neural Network Circuits", *IEEE Circuits and Devices magazine*, vol. 5, no. 4, p. 44, 1989.

41. W. Hubbard, D. Schwartz, J. Denker, H.P. Graf, R. Howard, L. Jackel, B. Straughn, and D. Tennant, "Electronic Neural Networks", *Neural Networks for Computing*, vol. 151, pp. 227-234, 1986.

42. H.P. Graf, L.D. Jackel, R.E. Howard, B. Straughn, J.S. Denker, W. Hubbard, D.M. Tennant, and D. Schwartz, "VLSI Implementation of a Neural Network Memory with Several Hundreds of Neurons", *AIP Conference Proceedings 151: Neural Networks for Computing*, pp. 182-187, 1986.

43. H.H. Busta, O.K. Ersoy, J.E. Pogemiller, K.D. Mackenzie, and R.W. Standley, "Hardware Implementation of a Wired-Once Neural Net in Thin-Film Technology on a Glass Substrate", *IEEE Transactions on Electron Devices*, vol. 37, no. 4, pp. 1039-1045, 1990.

44. M.D. Binns, F.J. Clough, and S.C.J. Garth, "An Architecture for Fully Integrated Large-Scale Neural Networks", *Proc. International Conference on Microelectronics for Neural Networks, Edinburgh*, pp. 187-194, 1993.

45. C.D. Kornfeld, R.C. Frye, C.C. Wong, and E.A. Rietman, "An Optically Programmed Neural Network", *Proc. IEEE International Conf. on Neural Networks 1988*, vol. 2, pp. 357-364.

46. U. Cilingiroglu, "A Purely Capacitive Synaptic Matrix for Fixed-Weight Neural Networks", *IEEE Transactions on Circuits and Systems*, vol. 38, no. 2, pp. 210-217, 1991.

47. A. P. Thakoor, J. L. Lamb, A. Moopenn, and J. Lambe, "Binary Synaptic Connections based on Memory Switching in a-Si:H", in *AIP Conference Proceedings 151, Neural Networks for Computing*, ed. John S. Denker, pp. 426 - 431, American Institute of Physics, 1986.

48. J.L Lamb, A.P. Thakoor, A. Moopenn, and S.K. Khanna, "Resistive Synaptic Interconnects for Electronic Neural Networks", *Journal of Vacuum Science and Technology A-Vacuum Surfaces and Films*, vol. 5, no. 4, pp. 1407-1411, 1987.

49. C.A. Mead, "Adaptive retina", in *Analog VLSI Implementation of Neural Systems*, ed. Carver Mead and Mohammed Ismail, pp. 239-246, Kluwer Academic Publishers, Boston, MA, 1989.

50. R.L. Sigvartsen, "An Analog Neural Network Chip with On-Chip Learning", *Masters Thesis, University of Oslo*, August 1994.

51. G. Cauwenberghs, C.F. Neugebauer, and A. Yariv, "An Adaptive CMOS Matrix-Vector Multiplier for Large Scale Analog Hardware Neural Network Applications", *Proc. IJCNN-91-Seattle*, vol. 2, pp. 507-511.

52. R. Tawel, R. Benson, and A.P. Thakoor, "A CMOS UV-Programmable Nonvolatile Synaptic Array", *Proc. IJCNN-91-SEATTLE*, vol. 2, pp. 581-585.

53. M. Holler, S. Tam, H. Castro, and R. Benson, "An Electrically Trainable Artificial Neural Network (ETANN) with 10240 Floating Gate Synapses", *Proc. International Joint Conference on Neural Networks*, 1989.

54. Intel Corporation, "80170NX Electrically Trainable Analog Neural Network", *ETANN datasheet*, June 1991.

55. A. Kramer, C.K. Sin, R. Chu, and P.K. Ko, "Compact EEPROM Based Weight Functions", *Proc. Advances in Neural Information Processing Systems 3*, pp. 1001-1007, 1990.

56. R.L. Shimbukuro, R.E. Reedy, and G.A. Garcia, "Dual-Polarity Nonvolatile MOS Analogue Memory (MAM) Cell for Neural-Type Circuitry", *Electronics Letters*, vol. 24, no. 19, pp. 1231-1232, 1988.

57. R.L. Shimbukuro, P.A. Shoemaker, and M.E. Stewart, "Circuitry for Artificial Neural Networks with Nonvolatile Analog Memories", *Proc. 1989 IEEE International Symp on Circuits and Systems (ISCAS-89)*, vol. 3, pp. 1217-1220.

58. R.S. Withers, R.W. Ralston, and E. Stern, "Nonvolatile Analog Memory in MNOS Capacitors", *IEEE Electron Device Letters*, vol. EDL1, no. 2, pp. 42-45, 1980.

59. J.P. Sage, R.S. Withers, and K.E. Thompson, "MNOS/CCD Circuits for Neural Network Implementations", *Proc. ISCAS 89*, pp. 1207-1209.

60. B. Widrow, *DARPA - NEURAL NETWORK STUDY*, pp. 601-619, AFCEA International Press, November 1988.

61. L.T. Clark, R.O. Grondin, and S.K. Dey, "Integrated-Circuit Neural Networks using Ferroelectric Analog Memory", *Proc. 11th International Phonenix Conference on Computers and Communications (IPCCC- 92)*, pp. 736-742.

62. R. Ramesham, S. Thakoor, T. Daud, A.P. Thakoor, and S.K. Khanna, "Thin-Film Solid-State Nonvolatile Electrochemically Programmable Reversible Resistor for Electronic Neural Networks", *Journal of the Electrochemical Society*, vol. 135, no. 3, p. 155C, 1988.

63. R. Ramesham, S. Thakoor, T. Daud, and A.P. Thakoor, "Solid-state Reprogrammable Analog Resistive Devices for Electronic Neural Networks.", *Journal of the Electrochemical Society*, vol. 137, no. 6, pp. 1935-1939, 1990.

64. E.G. Spencer, "Programmable Bistable Switches and Resistors for Neural Networks", *AIP Conference Proceedings - Snowbird 1986*, pp. 414-419, American Institute of Physics.

65. T. Triffet and H.S. Green, "Development of an Electrochemical Transistor for use as an Artificial Neural Synapse", *Proc. 3rd International Conference on Microelectronics for Neural Networks*, pp. 195-205.

66. H.C. Card and W.R. Moore, "EEPROM Synapses Exhibiting Pseudo-Hebbian Plasticity", *Electronics Letters*, vol. 25, no. 12, pp. 805-806, 1989.

67. M. Jabri and B. Flower, "Weight Perturbation - An Optimal Architecture and Learning Technique for Analog VLSI Feedforward and Recurrent Multilayer Networks", *IEEE Transactions on Neural Networks*, vol. 3, no. 1, pp. 154-157, 1992.

68. A.J. Montalvo, R.S. Gyurcsik, and J.J. Paulos, "Building Blocks for a Temperature Compensated Analog VLSI Neural Network with On-chip Learning", *Proceedings of the 1994 IEEE International Symposium on Circuits and Systems*, vol. 6, pp. 363-366.

69. J. Hajto and A.E. Owen, "Applications of Amorphous Semiconductor Materials in Electronics", *Materials & Design*, vol. 5, pp. 221-242, 1984.

70. A.E. Owen, P.G. LeComber, J.Hajto, M.J. Rose, and A.J. Snell AJ International Journal of Electronics, *Switching in Amorphous Devices*, 73, pp. 897-906, 1992.

71. J. Hajto, A.E. Owen, A.J. Snell, P.G. LeComber, and M.J. Rose, "Analog Memory and Ballistic Electron Effects in Metal-Amorphous Silicon Structures", *Philosophical Magazine B - Physics of Condensed Matter Structural Electrinic Optical and Magnetic Properties*, vol. 63, no. 1, pp. 34-369, 1991.

72. J. Hajto, A.E. Owen, S.M. Gage, A.J. Snell, P.G. LeComber, M.J. Rose, and Physical Review Letters, *Quantized Electron-Transport in Amorphous-Silicon Memory Structures*, 66, pp. 1918-1921, 1991.

73. M.J. Rose, J. Hajto, P.G. LeComber, S.M. Gage, W.K. Choi, A.J. Snell, and A.E. Owen, "Amorphous-silicon Analog Memory Elements", *Journal of Non-Crystalline Solids*, vol. 115, no. 1-3, pp. 168-170, 1989.

74. A.A. Reeder, I.P. Thomas, C. Smith, J. Wittgreffe, D.J. Godfrey, J. Hajto, A.E. Owen, A.J. Snell, A.F. Murray, M.J. Rose, I.S. Osbourne, and P.G. LeComber, "Application of Analogue a-Si Memory Devices to Resistive Synapses for Neural Networks", *Proc. Materials Research Society Symposium*, vol. 258, pp. 1081-1086, 1992.

75. M. Jafar and D. Haneman, "Switching in Amorphous-Silicon Devices", *Physical Review B - Condensed Matter*, vol. 49, no. 7, pp. 13611-13615, 1994.

76. A.A. Reeder, I.P. Thomas, C. Smith, J. Wittgreffe, and D.J. Godfrey, "Application of Analogue a-Si Memory Devices to Resistive Synapses for Neural Networks", *BT Technology Journal*, vol. 10, no. 3, pp. 155-160, July 1992.

77. M.J. Rose, J. Hajto, P.G. LeComber, A.J. Snell, A.E. Owen, and I.S. Osborne, "Amorphous-silicon Analog Memory Elements", *Materials Research Society Symposium Proceedings*, vol. 219, pp. 525-530, 1991.

78. A.F. Murray and A.V.W. Smith, "Asynchronous Arithmetic for VLSI Neural Systems", *Electronics Letters*, vol. 23, no. 12, pp. 642-643, 1987.

79. D.J. Baxter, "Process-Tolerant VLSI Neural Networks for Applications in Optimisation", *Ph.D. Thesis (University of Edinburgh)*, 1993.

80. S. Churcher, "VLSI Neural Networks for Computer Vision", *Ph.D. Thesis (University of Edinburgh)*, 1993.

81. J.A. McDonald, "Neural Nets are Starting to make Sense", *Biosensors and Bioelectronics*, vol. 7, no. 9, pp. 621-626, 1992.

82. L.M. Reyneri, M. Chiaberge, and D. Delcorso, "Using Coherent Pulsewidth and Edge Modulations in Artificial Neural Systems", *International Journal of Neural Systems*, vol. 4, no. 4, pp. 407-418, 1993.

83. M. Cao, T. Zhao, K.C. Saraswat, and J.D. Plummer, "A Simple EEPROM Cell Using Twin Polysilicon Thin Film Transistors", *IEEE Electron Device Letters*, vol. 15, no. 8, pp. 304-306, 1994.

84. C.K. Sin, A. Kramer, V. Hu, R.R. Chu, and P.K. Ko, "EEPROM as an Analog Storage Device, with Particular Applications in Neural Networks", *IEEE Transactions on Electron Devices*, vol. 39, no. 6, pp. 1410-1419, 1992.

85. A. Wright, "Analog Data Storage - Speaking of the Future", *Electronics World and Wireless World*, vol. 98, no. 1671, pp. 110-113, 1992.

86. E. Sackinger and W. Guggenbuhl, "An Analog Trimming Circuit Based on A Floating Gate Device", *IEEE Journal of Solid State Circuits*, vol. 23, no. 6, pp. 1437-1440, 1988.

87. J.L. Meador, A. Wu, C. Cole, N. Nintunze, and P. Chintrakulchai, "Programmable Impulse Neural Circuits", *IEEE Transactions on Neural Networks*, vol. 2, no. 1, pp. 101-109, 1991.

88. F. Devos, M. Zhang, Y. NI, and J.F. Pone, "Trimming CMOS Smart Imager with Tunnel-Effect Nonvolatile Analog Memory", *Electronics Letters*, vol. 29, no. 20, pp. 1766-1767, 1993.

89. A. Thomsen and M.A. Brooke, "A Floating Gate MOSFET with Tunnelling Injector Fabricated Using a Standard Double-Polysilicon Process", *IEEE Electron Device Letters*, vol. 12, no. 3, pp. 111-113, 1991.

90. L.R. Carley, "Trimming Analog Circuits using Floating-Gate Analog MOS Memory", *IEEE Journal of Solid State Circuits*, vol. 24, no. 6, pp. 1569-1575, 1989.

91. S. Kim, Y.C. Shin, N.C.R. Bogineni, and R. Sridhar, "A Programmable Analog CMOS Synapse for Neural Networks", *Analog Integrated Circuits and Signal Processing*, vol. 2, no. 4, pp. 345-352, 1992.

92. S.T. Wang, "On the IV Characteristics of Floating-Gate MOS Transistors", *IEEE Transactions on Electron Devices*, vol. ED-26, no. 9, pp. 1292-1294, 1979.

93. A. Kolodny, S. Nieh, B. Eitan, and J. Shappir, "Analysis and Modelling of Floating-Gate EEPROM Cells", *IEEE Transactions on Electron Devices*, vol. 33, no. 6, pp. 835-844, 1986.

94. T. Ong, P.K. Ko, and C. Hu, "The EEPROM as an Analog Memory Device", *IEEE Transactions on Electron Devices*, vol. 36, no. 9, pp. 1840-1841, 1989.

95. G. Vansteenwijk, K. Hoen, and H. Wallinga, "A Nonvolatile Analog Programmable Voltage Source Using the VIPMOS EEPROM structure", *IEEE Journal of Solid State Circuits*, vol. 28, no. 7, pp. 784-788, 1993.

96. T. Morie, "A Ballistic Analog Memory Device for Neural Network Implementation", *Solid State Electronics*, vol. 34, no. 8, pp. 919-920, 1991.

97. O. Fujita, Y. Amemiya, and A. Iwata, "Characteristics of Floating Gate Memory Device as Analogue Memory for Neural Networks", *Electronics Letters*, vol. 27, no. 11, pp. 924-926, 1991.

98. S. Bibyk and M. Ismail, "Neural network building blocks for analog MOS VLSI", in *Analogue IC Design: The Current Mode Approach*, ed. C. Toumazou, F.J. Lidgey and D.G. Haigh, pp. 600-605, London: Peter Peregrinus, 1990.

99. B.W. Lee, B.J. Sheu, and H. Yang, "Analog Floating-Gate Synapses for General Purpose VLSI Neural Computation", *IEEE Transactions on Circuits and Systems*, vol. 38, no. 6, pp. 654-658, 1991.

100. D.A. Durfee and F.S. Shoucair, "Comparison of Floating Gate Neural Network Memory Cells in Standard VLSI CMOS Technology", *IEEE Transactions on Neural Networks*, vol. 3, no. 3, pp. 347-353, 1992.

101. D.A. Durfee and F.S. Shoucair, "Low Programming Voltage Floating Gate Analog Memory Cells in Standard VLSI CMOS Technology", *Electronics Letters*, vol. 28, no. 10, pp. 925-927, 1992.

102. A.J. Montalvo and J.J. Paulos, "Improved Floating Gate Devices using Standard CMOS Technology", *IEEE Electron Device Letters*, vol. 14, no. 8, pp. 372-374, 1993.

Appendix G

List of Publications

References

1. A.F. Murray, S. Churcher, A. Hamilton, A.J. Holmes, G.B. Jackson, and R.J. Woodburn, "Pulse-Stream VLSI Neural Networks", *IEEE MICRO*, pp. 29-39, 1994.
2. A.J. Holmes, S. Churcher, J. Hajto, A.F. Murray, and M.J. Rose, "Pulsewidth Synapses incorporating a-Si:H Non-Volatile Analogue Memories", *Neural Information Processing Systems (NIPS) Conference*, 1994. In Press
3. A.J. Holmes, A.F. Murray, A.J. Snell, J. Hajto, M. Rose, and R. Gibson, "Design of Analogue Synapse Circuits using Non-Volatile a-Si:H Memory Devices", *Int. Symposium on Circuits and Systems, London*, pp. 351-354, 1994.
4. A.J. Holmes, R.A.G. Gibson, J. Hajto, A.F. Murray, A.E. Owen, M.J. Rose, and A.J. Snell, "Use of a-Si-H Memory Devices for Nonvolatile Weight Storage in Artificial Neural Networks", *Journal of Non-crystalline Solids*, vol. 166, pp. 817-820, June 1993.
5. J. Hajto, A.J. Snell, M.J. Rose, A.E. Owen, I.S. Osbourne, T. Kosa, A.J. Holmes, and Materials Research Symposium Proceedings, *Step-Like Current-Voltage Characteristics in Metal a-Si-H Metal Structures*, 297, pp. 1061-1066, 1993.
6. J. Hajto, A.J. Snell, A. Holmes, A.E. Owen, M.J. Rose, and R.A.G. Gibson, "DC and AC Measurements on Metal/a-Si-H/Metal Thin-Film Devices", *Journal of Non-crystalline Solids*, vol. 166, pp. 821-824, 1993.

Use of a-Si:H memory devices for non-volatile weight storage in artificial neural networks

A.J.Holmes[a] R.A.G.Gibson[b] J.Hajto[a], A.F.Murray[a], A.E.Owen[a], M.J.Rose[b] and A.J.Snell[a] *

[a]Dept. of Electrical Engineering,
University of Edinburgh, EH9 3JL, Scotland

[b]Dept. of Applied Physics and Electronic & Manufacturing Engineering,
University of Dundee, DD1 4HN, Scotland

An Artificial Neural Network (ANN) is an ensemble of simple processing units interconnected by variable strength weights. VLSI ANNs use either dynamic techniques or non-volatile EEPROM technology for weight storage. a-Si:H memory devices offer an alternative method for the non-volatile storage of analogue weight values. Results are presented from a test chip on which a-Si:H analogue memory devices were fabricated on the surface of a conventional CMOS chip. The design of a second chip is discussed, in which a-Si:H devices are used to store the synaptic weights.

1. Overview

A brief introduction to a-Si:H memory devices and Neural Networks is followed by a description of ANN devices which incorporate a-Si:H. The next section covers the design of and results from our first a-Si:H test chip. Finally the design of the latest chip, on which a-Si:H devices are used for synaptic weight storage, is discussed.

1.1. a-Si:H Memory Devices

The a-Si:H memory device [1] comprises a 1000Å thick layer of p^+ a-Si:H sandwiched between Vanadium and Chromium electrodes (Figure 1).

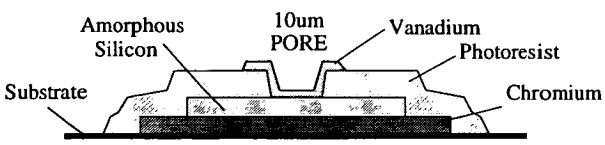


Figure 1. Construction of a-Si:H device

After fabrication the a-Si:H device has a very large (several GΩ) resistance, owing to the metal-semiconductor Schottky barriers at the contacts. To program the device into a lower resistance state the following steps must be carried out:-

- Forming: This is a once only process. A series of 300ns pulses, increasing in amplitude from 5v to 14v, is applied across the device electrodes. This creates a vertical conducting channel which can be programmed to a value in the range 1KΩ to 1MΩ.
- Write: To decrease the device's resistance, negative, "Write", pulses are applied.
- Erase: To increase the device's resistance, positive, "Erase", pulses are applied.
- Read: The device resistance can be *read* using a voltage of less than 0.5v without causing reprogramming.

The programming pulses, which range between 2v and 5v, are typically 120ns in duration. This means that the programming is potentially much faster than for EEPROM devices used in the same context, which use a series of 100μs pulses to set the threshold voltage [2].

1.2. Neural Networks

An Artificial Neural Network is a stylised model of its biological counterpart. An ANN is an ensemble of simple processing units connected together by variable strength weights. The basic

*Research sponsored by BT.

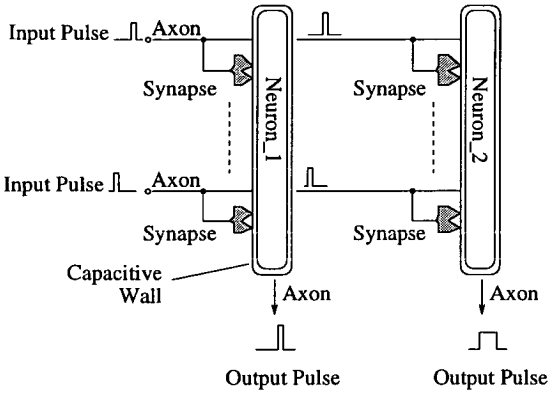


Figure 2. Model of Biological Neural Network

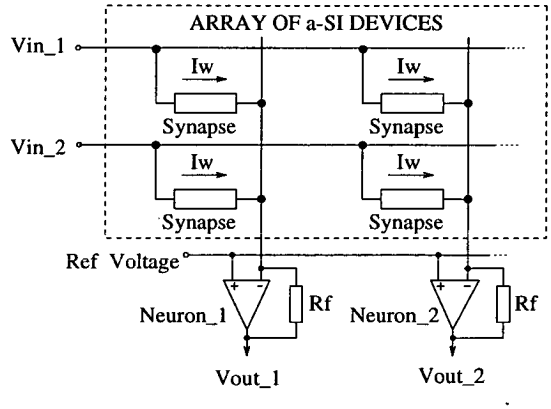


Figure 3. Original a-Si:H Neural Network

processing element is the neuron. This sums the large number of input signals and then outputs a current pulse once a voltage threshold level has been exceeded. The inputs are current pulses which are transmitted to the neuron through variable strength junctions called synapses, as shown in Figure 2. Synapses, which are either inhibitory or excitatory, can be thought of as a $\pm$ multiplication term. The function to be performed by an ANN can therefore be summarised as:

$$S_j = f(\sum_i W_{ji} S_i) \quad (1)$$

S_i = *Input State* W_{ji} = *Synaptic Weight*
S_j = *Output State* $f()$ = *Threshold function*

Our aim is to use the a-Si:H memory to store the analogue programmable weight W_{ji}.

1.3. Artificial Neural Networks and a-Si:H

The recent interest in Neural Networks has been accompanied by a plethora of hardware (VLSI) implementations of ANN.

In many of these implementations the synapse value is stored dynamically, that is as a voltage on a capacitor. This weight voltage must therefore be refreshed from off-chip digital RAM. There are also examples of chips that use EEPROM technology for non-volatile weight storage [2].

a-Si:H has been used in a number of hardware NNs either as high value resistors, in mask programmable arrays[3, 4], or as light dependent resistors, in neural networks with optical inputs[5]. An earlier design in this project [6] used a 10x10 array of a-Si:H memory devices, representing the synaptic weights, as the input resistors to a current summing op-amp, as shown in Figure 3.

However this architecture and computational style has a number of disadvantages. The input signal must be kept within the range from 0v to 0.5v, while the synaptic weights are limited to positive values. Furthermore, external op-amp circuitry is required.

2. Combining a-Si:H and CMOS

In the next design a-Si:H memories were integrated with the op-amp circuitry on a single CMOS chip.

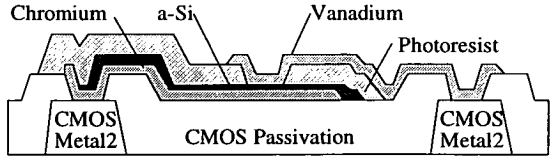


Figure 4. a-Si:H on CMOS backplane

This was achieved by fabricating the a-Si:H on the final passivated surface of the CMOS chip, connecting it to the Metal 2 layer through holes in the passivation, as shown in Figure 4.

With this approach, however, the substrate diodes must be protected, as their breakdown voltage is only 10v. The 14v pulse required during forming is therefore potentially damaging. To prevent damage, the address transistors were configured such that the top electrode of the device was maintained at a *safe* voltage level, as shown in Figure 5, while the pulse was applied across the a-Si:H device.

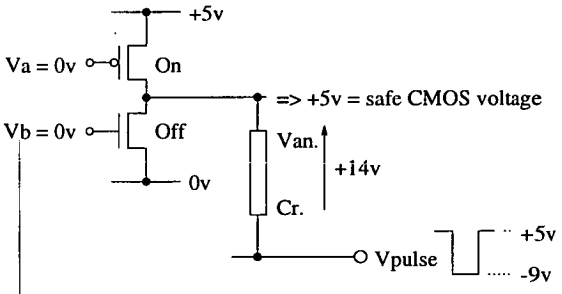


Figure 5. Circuit used for forming

A test chip, ASiTEST1, with a number of simple alternative programming circuits was designed and fabricated, to determine whether it was possible to fabricate and then program a-Si:H devices on a CMOS substrate.

2.1. ASiTEST1 - Results

Initial tests were performed on two terminal device structures (i.e. no address transistors). These were programmed successfully into a number of different resistance states, shown in Figure 6.

However, devices that incorporated access transistors displayed a much more limited range of resistance states. It was discovered that IV sweeps over a wider range (Figure 7) resulted in characteristics for the minimum resistance state which were stable up to about 3mA. Above this

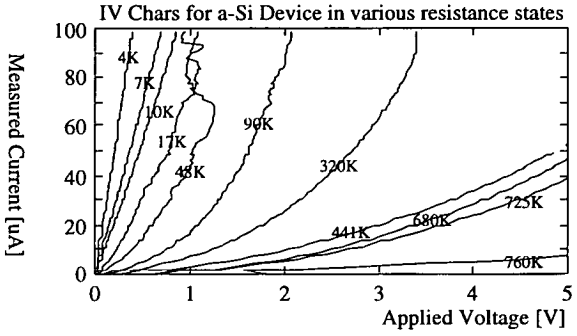


Figure 6. a-Si:H device in various resistance states

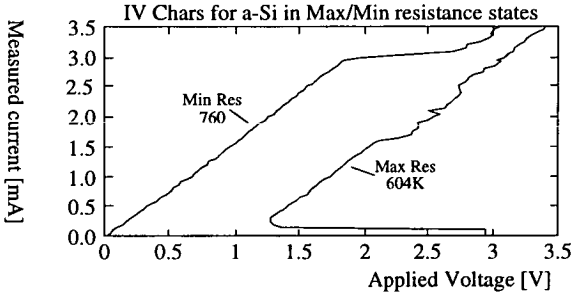


Figure 7. a-Si:H in Maximum and Minimum Resistance states

value, a transition occurs as the device changes state.

On the ASiTEST1 chip the Mosfets limited the maximum current through the a-Si:H to 1mA, which prevented the device reaching this transition region.

The latest design uses larger transistors which will hopefully allow the addressed device to be programmed successfully.

3. Chip with a-Si:H synapses

This chip uses a-Si:H for synaptic weight storage in a simple neural test circuit. Voltage levels are no longer used as the input signal, as in Figure 3. The input signals are provided as digital

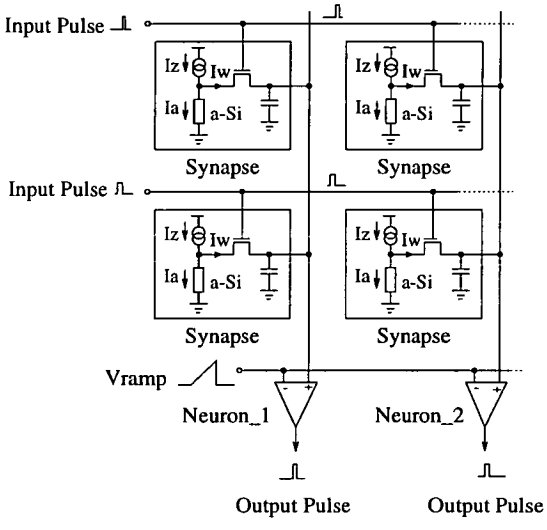


Figure 8. a-Si Pulse-stream Neural Network

pulses, with analogue information coded in the time domain - this is the Pulse Stream method [7]: In ASiTEST2, the width of the digital pulse codes the neural input. Digital signals are more noise tolerant and allow simpler inter chip communication than do purely analogue signals.

The circuit designed, shown in Figure 8, uses the a-Si:H memory to "store a weight current". This current is subtracted from a *zero* current to produce both positive (excitatory) and negative (inhibitory) values. This current is gated by the variable width input pulse which causes a packet of charge to be dumped on the neuron's integration capacitor.

Note: This capacitance is actually distributed amongst all the synapses. This means that the neuron does not have to be altered if more synapses are added: The extra capacitance needed to maintain the same output range is included in the additional cells.

The neuron is based on a design used in earlier dynamic storage ANN [8]. When the voltage on the neuron's integration capacitor exceeds the threshold level the output becomes a binary '1'. The threshold signal is normally a sigmoid or ramp waveform.

The chip ASiTEST2, which contains a number of different synapse test circuits has been designed and is currently being fabricated.

4. Conclusions

We have shown that a-Si:H analogue memory devices can be successfully integrated with conventional CMOS circuitry. A synapse has been designed which uses a-Si:H for the non-volatile storage of analogue synaptic weights.

REFERENCES

1. M.J.Rose, J.Hajto, P.G.Lecomber, S.M.Gage, W.K.Choi, A.J.Snell and A.E.Owen, Journal of Non-Crystalline Solids, 115 (1989) 168.
2. M.Holler, S.Tam, H.Castro and R.Benson, Int Conf on N.N.s Proc, (1989) 191.
3. A.P.Thakoor, J.L.Lamb, A.Moopenn and J.Lambe, AIP Conf Proc 151, (1986) 426.
4. H.P.Graf, L.D.Jackel, R.E.Howard, B.Straughn, J.S.Denker, W.Hubbard, D.M.Tennant and D.Schwartz, AIP Conf Proc 151, (1986) 182.
5. M.D.Binns, F.J.Clough and S.C.J. Garth, Int Conf on Microelectronics for N.Ns Proc, (1993) 182.
6. A.A.Reeder, I.P.Thomas, C.Smith, J.Wittgreffe, D.Godfrey, J.Hajto, A.Owen, A.J.Snell, A.F.Murray, M.Rose, I.S.Osborne and P.G.LeComber, Int Conf N.N.s Proc, (1991) 253.
7. A.F.Murray, A.Hamilton, D.J.Baxter, S.Churcher, H.M.Reekie and L.Tarassenko, IEEE Trans Neural Networks, (1993) 385.
8. S.Churcher, D.J.Baxter, A.Hamilton, A.F.Murray and H.M.Reekie, Proc NIPS5, (1991) 773.

Design of Analogue Synapse Circuits using Non-Volatile a-Si:H Memory Devices

A.J.Holmes, S.Churcher, J.Hajto & A.F.Murray
Department of Electrical Engineering
University of Edinburgh
King's Buildings,
Mayfield Road,
Edinburgh EH9 3JL
email ajho@ee.ed.ac.uk

M.J.Rose
Department of Applied Physics and Electronics
University of Dundee
Dundee DD1 4HN

INTRODUCTION

Analogue hardware implementations of neural networks have hitherto been hampered by the lack of a straightforward analogue memory capability. The synaptic weights which are developed by the network learning process must be stored (preferably at each synapse site) in order that a network can adequately perform any recall or classification tasks. Ideally, the storage mechanism should be compact, non-volatile, easily reprogrammable, and simple to implement.

Techniques which have been used to date include resistors (these are not generally reprogrammable, and suffer from being large and difficult to fabricate with any accuracy), dynamic capacitive storage [1] (this is compact, reprogrammable and simple, but implies an increase in system complexity, arising from off-chip refresh circuitry), EEPROM ("floating gate") memory [2] (which is compact, reprogrammable, and non-volatile, but cannot be reprogrammed in situ), and local digital storage (which is non-volatile, easily programmable and simple, but is very costly in "real estate" terms).

In this paper, we demonstrate that novel amorphous silicon memory devices can be incorporated into standard CMOS synapse circuits, to provide an analogue weight storage mechanism which is compact, non-volatile, easily reprogrammable, and simple to implement.

THE EPSILON SYNAPSE

Dynamic synapse circuits were used in the construction of a 120 input, 30 output chip christened EPSILON [1]. The fundamental function to be performed in an ANN is:

$$S_j = f(\sum_i W_{ji} S_i) \qquad (1)$$

$S_i = Input\ State$ $\quad W_{ji} = Synaptic\ Weight$
$S_j = Output\ State$ $\quad f() = Threshold\ function$

On the EPSILON chip the input and output states are represented in the form of pulse-streams [3]. A pulse-stream signal is a digital waveform within which the neural state or activity is encoded as the width or frequency of the on-state pulses. With Pulse Stream arithmetic, very simple two-quadrant synapse multiplier circuits can be designed, as shown in Figure 1.

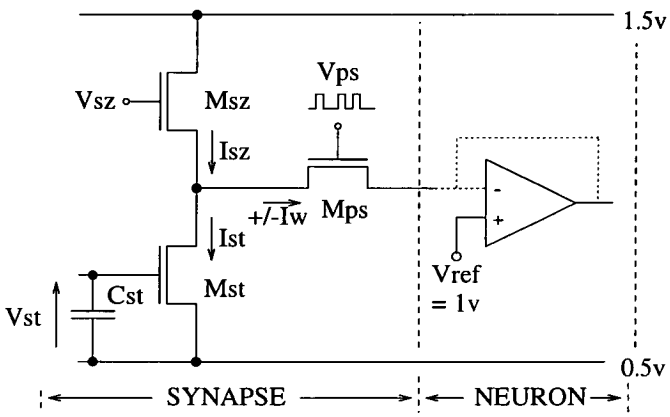


Figure 1: The EPSILON Synapse

By keeping Vds constant we ensure that the transistors Msz and Mst both operate in the linear regime. This ensures that Ids is directly proportional to Vgs. The stored current, Ist, is subtracted from the synapse zero current, Isz, to give a +/- weight current, Iw. This current is then gated by the variable width pulsestream sig-

nal. The packets of charge from each synapse are then integrated by the neuron circuitry.

As the figure shows, the weight voltage is stored as a voltage on a capacitor. Voltage decay with time necessitates external refresh circuitry. The complexity of the system would therefore be reduced if we could store the synaptic weights on-chip, in non-volatile memory devices. In this paper we look at the design of synapse cells that use novel a-Si:H analogue memories to provide on-chip, non-volatile weight storage.

a-Si:H MEMORY DEVICES

The a-Si:H analogue memory device [4] comprises a 1000Å thick layer of p^+ a-Si:H sandwiched between Vanadium and Chromium electrodes (Figure 2).

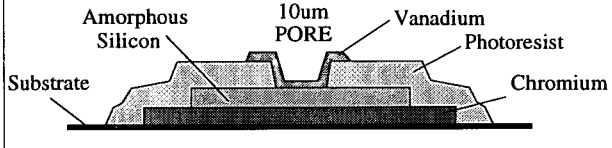


Figure 2: Construction of a-Si:H device

The a-Si device takes the form of a two-terminal, programmable resistor. It is an "add-on" to a conventional CMOS process, and does not demand that the normal CMOS fabrication cycle be disrupted. The a-Si device sits on top of the completed chip circuitry, making contact with the CMOS arithmetic elements via holes cut in the passivation layer. After fabrication a number of steps must be carried out in order to program the device to a given resistance state.

Programming

Before the a-Si device is usable, the following steps must be carried out:

- Forming: This is a once only process, applied to the a-Si device in its "virgin" state, where it has a resistance of several MΩ. A series of 300ns pulses, increasing in amplitude from 5v to 14v, is applied across the device electrodes. This creates a vertical conducting channel which can be programmed to a value in the range 1KΩ to 1MΩ.
- Write: To decrease the device's resistance, negative, "Write", pulses are applied.
- Erase: To increase the device's resistance, positive, "Erase", pulses are applied.
- Read: Pulses below 0.5v do not change the device resistance. The resistance can therefore be *read* using a voltage of less than 0.5v without causing reprogramming.

Programming pulses, which range between 2v and 5v, are typically 120ns in duration. Programming is therefore potentially much faster than for other EEPROM (floating gate) devices used in the same context, which use a series of 100μs pulses to set the threshold voltage [2].

ASiTEST1 Chip

To utilise existing neural circuitry, the a-Si:H memories were fabricated on the surface of a conventional CMOS wafer, as shown in Figure 3.

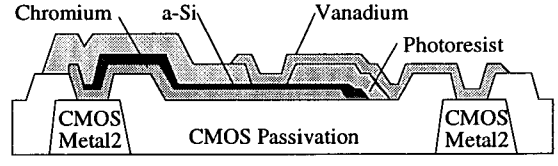


Figure 3: a-Si:H memory on a CMOS substrate

A test chip, ASiTEST1, was designed and fabricated [5]. This chip contains both CMOS addressed and discrete, two terminal, a-Si:H memory devices. Figure 4 shows a set of IV curves from one of the two terminal devices.

ASiTEST1 proved that working a-Si:H memories could be fabricated on CMOS substrates. The next sections describes synapse circuits using the a-Si:H devices. These new synapses use the reprogrammable a-Si:H resistor in the place of a storage capacitor or EEPROM cell. The following sections look at three different approaches to the incorporation of a-Si:H memory to store synaptic weights.

GLOBAL CURRENT MIRROR SYNAPSE

In this first design the a-Si:H device is driven by a globally distributed current mirror. The mirrored current produces a voltage drop that is equivalent to the voltage that was previously stored using a capacitor, as determined by the a-Si resistance value.

The circuit can be analysed by overlaying the current mirror characteristic on the a-Si:H data as shown in Figure 6. With Iset = 25uA the range of Vst is 1v to 5v.

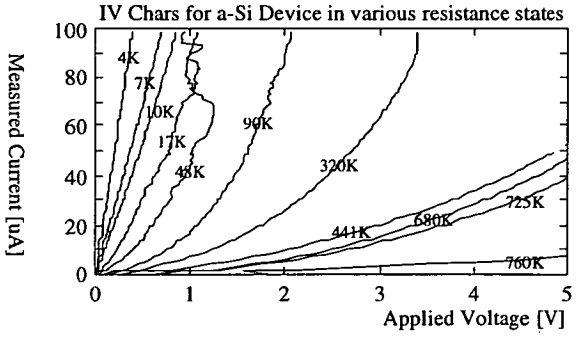


Figure 4: a-Si:H Memory device in various resistance states

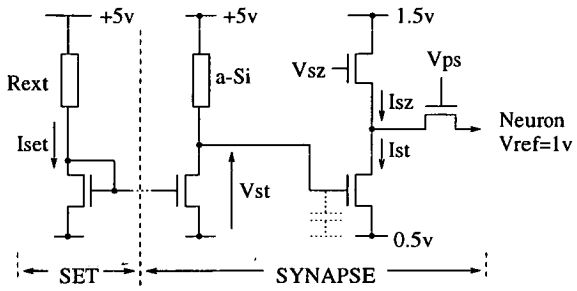


Figure 5: Synapse with Global Current Mirror

In order to minimise the effects of process variation, such as mirror mismatch, we do not monitor the resistance of the a-Si device during programming, rather we monitor the neuron output level for the chosen synapse. The device resistance can then be adjusted to give the desired Iw value rather than a particular resistance.

In this synapse the a-Si has effectively been used to store a weight voltage, the next two designs use the a-Si to store a weight current.

CONSTANT 0.5V SYNAPSE

Earlier designs using a-Si:H memory devices [6] showed that the operating voltage had to be below 0.5v to prevent reprogramming. However, results from the ASiTEST1 chip suggest that as long as either the voltage is kept below 0.5v or the current is kept below 50uA then the device will not change resistance. While synapses 1 and 3 operate below the current threshold it was decided that we should also design a synapse that operated below this 0.5v limit.

Since transistors in the EPSILON synapses operate with a constant Vds of 0.5v, one possible solution would be

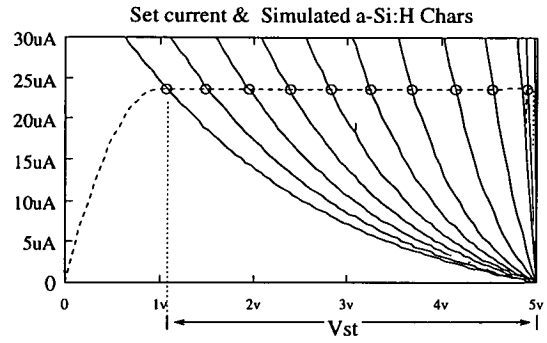


Figure 6: Global Current Mirror Synapse - Load Line

to simply connect the a-Si:H between the Vref and 0.5v rails.

One drawback of this approach however is the fact that the negative weights are confined to a fairly narrow band of a-Si:H resistances i.e. where Iw is greater than Isz. To increase the exploitation of the available dynamic range an additional bias transistor is added as shown in Figure 7.

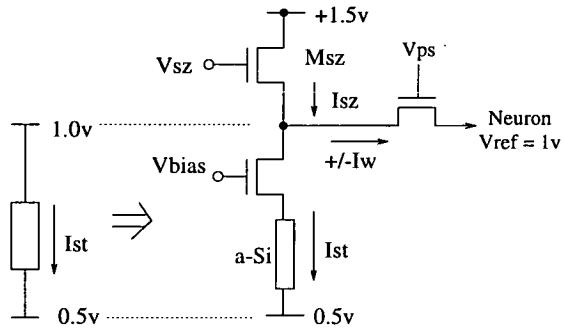


Figure 7: Constant 0.5v synapse

The effect of altering Vbias is illustrated in Figure 8. When Vbias is equal to +5v the transistor load line only intersects with one of the sample a-Si chars in the -ve Iw region, this is equivalent to the original case with no bias transistor. However, when we reduce Vbias to 2v the transistor has a higher Rds and so we intersect with more of the a-Si chars, which should hopefully mean that it will be easier to program the device to set a particular value of Iw.

ACTIVE RESISTOR SYNAPSE

In this design the complexity of the synapse is reduced by eliminating the external bias voltage required in the

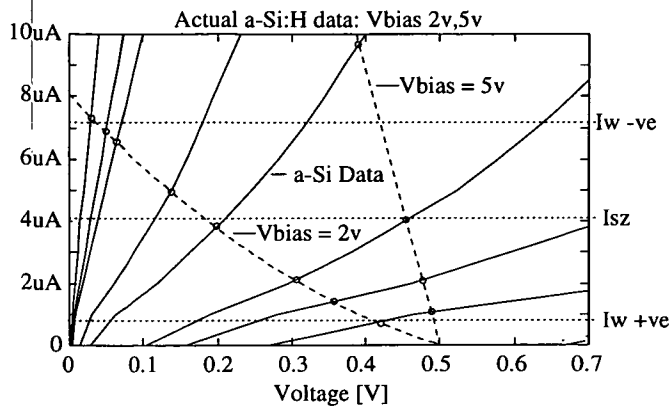


Figure 8: Constant 0.5v synapse

two earlier designs. An active resistor is placed in series with a a-Si:H memory to set a quiescent current which is then subtracted from Isz.

As with the other synapses the a-Si:H will be programmed to produce a given value of Iw, rather than a particular resistance state. This should hopefully minimize the effect of process variation on the active resistor characteristic.

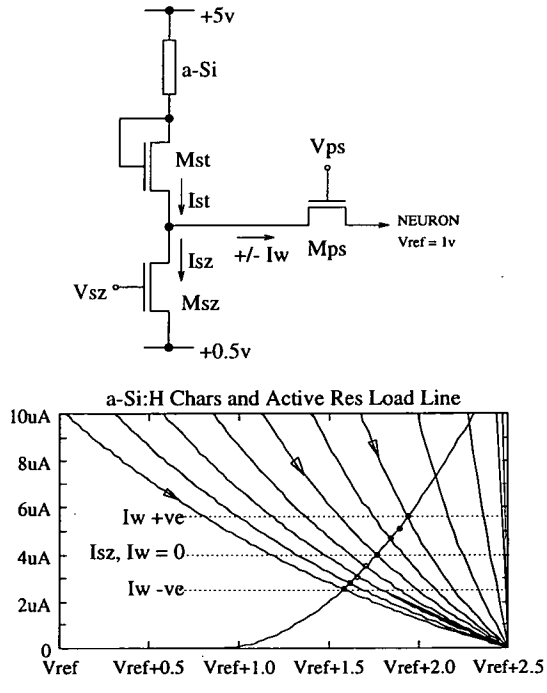


Figure 9: Synapse with Active Resistor

ASiTEST2 CHIP

In order to test these synapses, as well as other designs, a test chip, ASiTEST2, was designed.

This chip has been fabricated using the ES2 1.5um process. The operation of the circuits has been verified using external carbon resistors. We are currently awaiting the deposition of the a-Si:H layers prior to full testing.

CONCLUSION

We have shown that novel a-Si:H memory devices can be used to provide non-volatile weight storage in analogue VLSI neural networks. Future work is aimed at improving the yield and electrical properties of these devices. Networks based on these devices will be free from the complexities of refresh circuitry, and will therefore be well-suited to applications in the form of compact embedded neural systems, where the justification is strongest for analogue techniques in the neural context.

ACKNOWLEDGEMENTS

The authors would like to thank the U.K. Science and Engineering Research Council and BT for sponsoring this research.

References

[1] A.Hamilton, A.F.Murray, D.J.Baxter, S.Churcher and H.M.Reekie, Integrated Pulse-Stream Neural Networks, IEEE Trans on NN, pp385-393, 1992

[2] M.Holler, S.Tam, H.Castro and R.Benson, An Electrically Trainable ANN with 10240 Floating Gate Synapses, Int Conf on N.N.s Proc, pp191-196, 1989.

[3] A.F.Murray and A.V.W.Smith, Asynchronous Arithmetic for VLSI Neural Systems,Electronics Letters, vol 23, no 12, pp 642-643, June 1987.

[4] M.J.Rose, J.Hajto, P.G.Lecomber, S.M.Gage, W.K.Choi, A.J.Snell and A.E.Owen, Amorphous Silicon Analogue Memory Devices, Journal of Non-Crystalline Solids, vol 115, pp168-170, 1989.

[5] A.J.Holmes, R.A.G.Gibson, J.Hajto, A.F.Murray, A.E.Owen, M.J.Rose and A.J.Snell, Use of a-Si:H Memory Devices for Non-volatile Weight Storage in Artificial NNs, Proc ICAS 15, 1993

[6] A.A.Reeder et al., Application of Analogue Silicon Memory Devices to Resistive Synapse for NNs, MRS Symposium Proc, Vol 258, pp 1081-1085, 1992

Pulsestream Synapses with Non-Volatile Analogue Amorphous-Silicon Memories.

A.J. Holmes, A.F. Murray, S. Churcher and J. Hajto
Department of Electrical Engineering
University of Edinburgh
Edinburgh, EH9 3JL

M. J. Rose
Dept. of Applied Physics and Electronics,
Dundee University
Dundee DD1 4HN

Abstract

A novel two-terminal device, consisting of a thin 1000Å layer of p^+ a-Si:H sandwiched between Vanadium and Chromium electrodes, exhibits a non-volatile, analogue memory action. A circuit has been designed in which this device stores synaptic weights in an ANN chip, replacing the capacitor previously used for dynamic weight storage. Two different synapse designs are discussed and results are presented.

1 INTRODUCTION

Analogue hardware implementations of neural networks have hitherto been hampered by the lack of a straightforward (local) analogue memory capability. The ideal storage mechanism will be compact, non-volatile, easily reprogrammable, and will not interfere with the normal silicon chip fabrication process.

Techniques which have been used to date include resistors (these are not generally reprogrammable, and suffer from being large and difficult to fabricate with any accuracy), dynamic capacitive storage [4] (this is compact, reprogrammable and simple,

but implies an increase in system complexity, arising from off-chip refresh circuitry), EEPROM ("floating gate") memory [5] (which is compact, reprogrammable, and non-volatile, but is slow, and cannot be reprogrammed in situ), and local digital storage (which is non-volatile, easily programmable and simple, but consumes area horribly).

Amorphous silicon has been used for synaptic weight storage [1, 2], but only as either a high-resistance fixed weight medium or a binary memory.

In this paper, we demonstrate that novel amorphous silicon memory devices can be incorporated into standard CMOS synapse circuits, to provide an analogue weight storage mechanism which is compact, non-volatile, easily reprogrammable, and simple to implement.

2 a-Si:H MEMORY DEVICES

The a-Si:H analogue memory device [3] comprises a 1000Å thick layer of p^+ a-Si:H sandwiched between Vanadium and Chromium electrodes.

The a-Si device takes the form of a two-terminal, programmable resistor. It is an "add-on" to a conventional CMOS process, and does not demand that the normal CMOS fabrication cycle be disrupted. The a-Si device sits on top of the completed chip circuitry, making contact with the CMOS arithmetic elements via holes cut in the protective passivation layer, as shown in Figure 1.

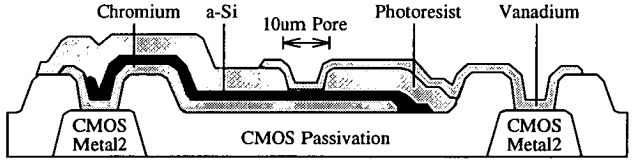


Figure 1: The construction of a-Si:H Devices on a CMOS chip

After fabrication a number of steps must be carried out in order to program the device to a given resistance state.

Programming, and Pre-Programming Procedures

Before the a-Si device is usable, the following steps must be carried out:

- Forming: This is a once-only process, applied to the a-Si device in its "virgin" state, where it has a resistance of several MΩ. A series of 300ns pulses, increasing in amplitude from 5v to 14v, is applied to the device electrodes. This creates a vertical conducting channel or filament which can subsequently be programmed to a value in the range 1KΩ to 1MΩ.
- Write: To decrease the device's resistance, negative "Write", pulses are applied.

- Erase: To increase the device's resistance, positive "Erase", pulses are applied.
- Read: Pulses below 0.5v do not change the device resistance. The resistance can therefore be *read* or utilised as a weight storage medium using a voltage of less than 0.5v without causing reprogramming.

Programming pulses, which range between 2v and 5v, are typically 120ns in duration. Programming is therefore much faster than for other EEPROM (floating gate) devices used in the same context, which use a series of $100\mu s$ pulses to set the threshold voltage [5].

The following sections describe synapse circuits using the a-Si:H devices. These synapses use the reprogrammable a-Si:H resistor in the place of a storage capacitor or EEPROM cell. These new synapses were implemented on a chip referred to as ASiTEST2, consisting of five main test blocks, each consisting of four synapses connected to a single neuron.

3 The EPSILON based synapse

The first synapse to be designed used the a-Si:H resistor as a direct replacement for the storage capacitor used in the EPSILON [4] synapse.

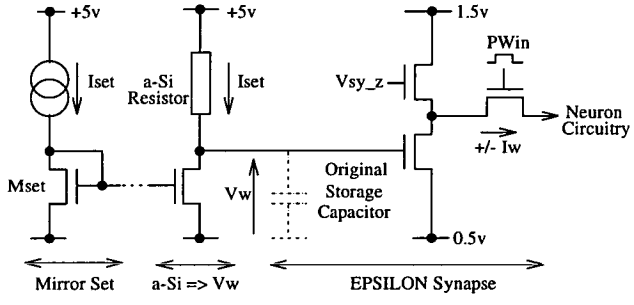


Figure 2: The EPSILON Synapse with a-Si:H weight storage

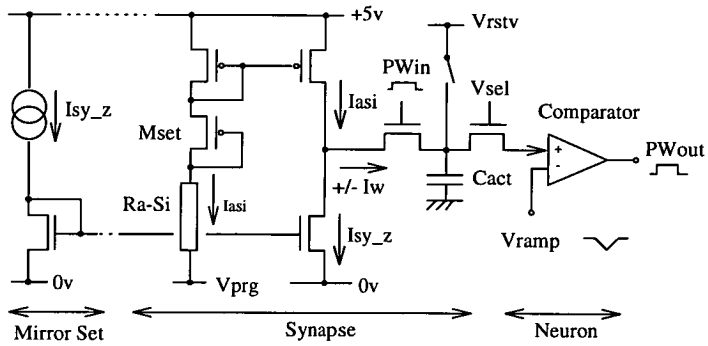
In the original EPSILON chip the weight voltage was stored as a voltage on a capacitor. In this new synapse design, shown in Figure 2, the a-Si:H resistance is set such that the voltage drop produced by Iset is equivalent to the original weight voltage, Vw, that was stored dynamically on the capacitor.

A new, simpler, synapse, which can be operated from a single +5v supply, was also be included on the ASiTEST2 chip.

4 The MkII synapse

The circuit is shown in Figure 3. The a-Si:H memory is used to store a current, Iasi. This current is subtracted from a zero current, Isy_z, to give a weight current , +/-Iw, which adds or subtracts charge from the activity capacitor, Cact.

For the circuit to function correctly we must limit the voltage on the activity capacitor to the range [1.5v,3.5v], this ensures that the transistors mirroring Isy_z and Iasi remain in saturation. As Figure 3 shows, there are few reference signals and the circuit operates from a single +5v power supply rail.



Power Supplies	References	Tail Currents
V5_0 = 5.0v	Vrstv = 2.5v Isy_z = 5uA	Ineu = 4uA

Figure 3: The MkII synapse

On first inspection the main drawback of this design would appear to be a reliance on the accuracy with which the zero current Isy_z is mirrored across an entire chip. The variation in this current means that two cells with the same synapse resistance could produce widely differing values of Iw. However, during programming we do not use the resistance of the a-Si:H device as a goal - rather we monitor the voltage on Cact for a given PWin signal. We then increase/decrease the resistance of the a-Si:H device until the desired voltage level is achieved.

Example: To set a weight to be the maximum positive value, we adjust the a-Si resistance until a PWin signal of 5us, the maximum input signal, gives a voltage of 3.5v on the integration capacitor.

We are able to set the synapse weight using the whole integration range of [1.5v,3.5v] by only closing Vsel for the desired synapse during programming. In normal operating mode all four Vsel switches will be closed so that the integration charge is summed over all four local capacitors.

4.1 Example - Stability Test

As an example of the use of integration voltage as means of monitoring the resistance of a particular synapse we have included a stability test. This was carried out on one of the test chips which contained the MkII synapse.

The four synapses on the test chip were programmed to give different levels of activation. The chip was then powered up for 30mins each subsequent day and the activation levels for each synapse were measured three times.

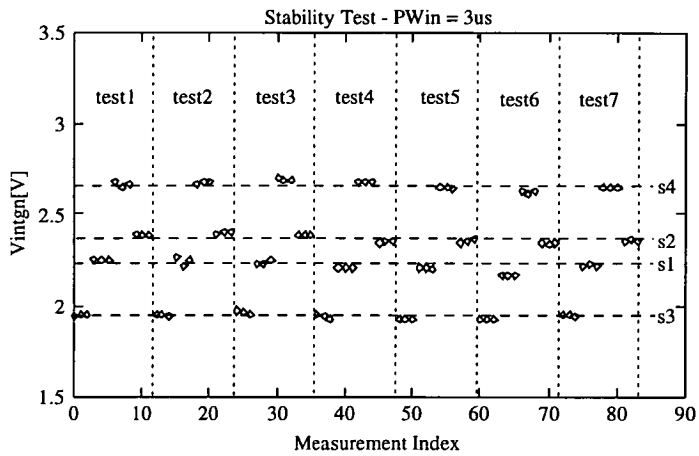


Figure 4: ASiTEST2- Stability Test

As figure 4 shows, the memories remain in the same resistance state (i.e retain their programmed weight value) over the whole 7-day period. Separate experiments on isolated devices indicate much longer hold times - of the order of months at least.

5 ASiTEST3

Recently we have received our latest, overtly neural, a-Si:H based test chip. This contains an 8x8 array of the MkII synapses.

The circuit board for this device has been constructed and partially tested while the ASiTEST3 chips are awaiting the deposition of the a-Si:H layers. We have been able to use an ASiTEST2 chip containing two of the MkII synapse test blocks i.e. 8 synapses and 2 neurons to exercise much of the board's functionality.

The test board contains a simple state machine which has four different states:

- State 0: Load Input Pulsewidths into SRAM from PC.
- State 1: Apply Input Pulsewidth signals to chip1.
- State 2: Use Vramp to generate threshold function for chip1. The resulting Pulsewidth outputs are used as the inputs to chip2, as well as being stored in SRAM.
- State 3: Use Vramp to generate threshold function for chip2. Read resulting Pulsewidth Outputs into SRAM.
- State 0: Read Output Pulsewidths from SRAM into PC.

The results obtained during a typical test cycle are shown in Figure 5.

As this figure shows different ramp signals, corresponding to different threshold functions, can be applied to chip1 and chip2 neurons.

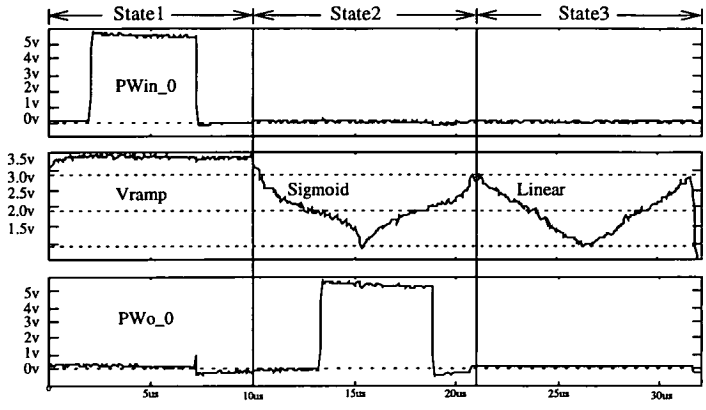


Figure 5: ASiTEST3 Board Scope Waveforms

While the signals shown in Figure 5 appear noisy the multiplier characteristic that the chip produces is still admirably linear, as shown in Figure 6. In this experiment all eight synapses on a test chip were programmed into different resistance states and PWin was swept from 0 to 3us.

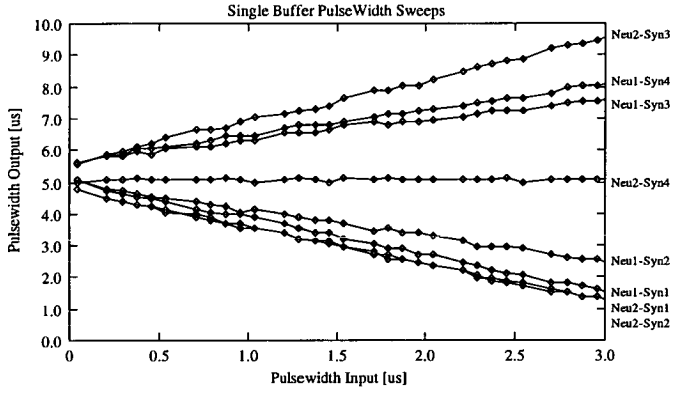


Figure 6: ASiTEST3 Board - MkII Synapse Characteristic

6 Conclusions

We have demonstrated the use of a-Si:H memory devices as a means of storing synaptic weights in a Pulsewidth ANN. We have also demonstrated the operation of an interface board which allows two 8x8 ANN chips, operating as a two layer network, to be controlled by a simple PC interface card.

This technology is most suitable for small networks in, for example, remote control

applications where cost and power considerations would favour a single all inclusive ANN chip with non-volatile, but programmable weights.

Another possible application of this technology will be in large networks constructed using thin film technology. If TFT's were used in place of the CMOS transistors then the area constraint imposed by crystalline silicon would be removed, allowing massively parallel networks to be integrated.

Acknowledgements

This research has been jointly funded by BT, formerly British Telecom, and SERC, the Science and Engineering Research Council.

References

[1] W. Hubbard et al.(1986) Electronic Neural Networks *AIP Conference Proceedings - Snowbird 1986* :227-234

[2] H.P. Graf (1986) VLSI Implementation of a NN memory with several hundreds of neurons *AIP Conference Proceedings - Snowbird 1986* :182-187.

[3] M.J. Rose et al (1989) Amorphous Silicon Analogue Memory Devices *Journal of Non-Crystalline Solids* **1**(115):168-170

[4] A.Hamilton et al. (1992) Integrated Pulse-Stream Neural Networks - Results, Issues and Pointers *IEEE Transactions on N.N.s* **3**(3):385-393

[5] M.Holler, S.Tam, H.Castro and R.Benson (1989) An Electrically Trainable ANN with 10240 Floating Gate Synapses. *Int Conf on N.N.s Proc* :191-196

[6] A.F.Murray and A.V.W.Smith.(1987) Asynchronous Arithmetic for VLSI Neural Systems. *Electronics Letters* **23**(12):642-643

[7] A.J. Holmes et al. (1993) Use of a-Si:H Memory Devices for Non-volatile Weight Storage in ANNs. *Proc ICAS 15* :817-820